UNIVERSIDADE FEDERAL DE JUIZ DE FORA
CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

MIGUEL NASCIMENTO ALMEIDA

ANÁLISE DO PORTFÓLIO CIENTÍFICO DA UFJF COM EMPREGO DE *TEXT MINING* EM DADOS DA WEB OF SCIENCE

JUIZ DE FORA
2021

MIGUEL NASCIMENTO ALMEIDA

ANÁLISE DO PORTFÓLIO CIENTÍFICO DA UFJF COM EMPREGO DE *TEXT MINING* EM DADOS DA WEB OF SCIENCE

Trabalho de Conclusão de Curso apresentado a Faculdade de Engenharia da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do título de Engenheiro de Produção.

Orientadora: DSc. Mariana Paes da Fonseca

JUIZ DE FORA

2021

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Almeida, Miguel .
LEVANTAMENTO DO PORTFÓLIO CIENTÍFICO DA UFJF COM O EMPREGO DE DATA MINING, TEXT MINING E DADOS DO LATTES / Miguel Almeida. -- 2021.
86 p.

Orientadora: Mariana Fonseca
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Faculdade de Engenharia, 2021.

1. Text Mining. 2. Data Mining. 3. Data Science. 4. Governança de Conhecimento. I. Fonseca, Mariana, orient. II. Título.

Miguel Nascimento Almeida

LEVANTAMENTO DO PORTFÓLIO CIENTÍFICO DA UFJF COM O EMPREGO DE *DATA MINING, TEXT MINING* E DADOS DO *LATTES*

Trabalho de Conclusão de Curso apresentado a Faculdade de Engenharia da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do título de Engenheiro de Produção.

Aprovada em 11 de março de 2021.

BANCA EXAMINADORA

D. Sc, Mariana Paes da Fonseca (Orientadora)

Universidade Federal de Juiz de Fora

p/ D. Sc, Clarice Breviglieri Porto

Universidade Federal de Juiz de Fora

p/ D. Sc, Roberta Cavalcanti Pereira Nunes

Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Agradeço ao universo, “fluxo”, ou ao encadeamento de situações que tornam possível o mundo ser belo da maneira que é. Independente de ser aleatório ou não – o que particularmente não acredito que seja – sou grato pelas oportunidades que me levaram à conclusão deste trabalho.

Grato pela família, em especial pais e irmã. Os primeiros que sempre enxergaram na educação a maior oportunidade de transformação e fizeram de tudo para que eu pudesse me transformar. E à minha irmã querida, exemplo de intelectualidade e dedicação acadêmica (sim, a família tem o orgulho de ter uma doutora).

Agradeço, com todo o coração, à minha orientadora Mariana: a “funcionária pública que todo contribuinte deseja”. Iniciamos o trabalho em meio a uma pandemia quando as atividades obrigatórias estavam suspensas, mesmo assim, por opção pessoal, escolheu me ajudar e seguirmos com os estudos de modo remoto e fazendo valer nosso tempo.

Agradeço à UFJF por oferecer educação de primeira qualidade e de modo gratuito, além de cada um dos autores citados durante essa revisão e aos professores convidados para compor a banca que tanto contribuíram positivamente para meu crescimento.

Finalizando, agradeço a oportunidade de ser humano, já que a capacidade de aprender e sorrir são prerrogativas exclusivas de nossa espécie.

Obrigado, obrigado, obrigado!

RESUMO

Está em voga afirmar que a era atual é a era dos dados, entretanto, cabe discernir o papel que dados, informação, conhecimento e sabedoria desempenham. Ao longo dos anos fica cada vez mais evidente que o desafio na construção de conhecimento deixou de ser a falta de dados e sim o oposto, seu volume colossal – o chamado *big data*. Considerando que a produção de conhecimento formal acontece majoritariamente nas Instituições de Ensino Superior, hoje se entende com clareza que tal produção, para ocorrer de modo eficiente, carece de atuação coordenada de fatores sistêmicos, coordenação essa que pode ser abordada pela Governança do Conhecimento. O presente trabalho, através de análise exploratória e aplicação de ferramentas de Ciência de Dados como *Data* e *Text Mining*, busca evidenciar informações que, de forma geral, existiam de maneira tácita para UFJF. Este estudo ambiciona oferecer ideia de caminho para o qual as pesquisas da instituição estão indo, além de ser capaz de fornecer análises qualitativas e quantitativas sobre o cenário atual da Instituição de Ensino explorada. Ademais, por meio do emprego de ferramentas estatísticas e computacionais, indica as áreas e que a UFJF é referência, quais temas foram os mais publicados ao longo dos últimos vinte anos, quais áreas apresentam maiores volumes de publicações e expõe, em formato de nuvens de palavras, aquelas que foram classificadas como mais relevantes por algoritmos de *machine learning*. A abordagem metodológica, após extração dos dados através da *Web of Science*, dividiu-se em duas frentes de atuação: uma em Excel e outra em Python. Na primeira ocorreu limpeza e análise exploratória dos dados, enquanto na segunda a manipulação foi mais complexa, envolvendo aplicação de *data mining*, *text mining*, *machine learning* e nuvens de palavras. Com a conclusão dessas etapas, foi possível constatar que as áreas da UFJF que contribuem com maiores números de publicação são as de Física, Química e Engenharia Elétrica. Ademais, pode-se perceber que existe uma distribuição de contribuição razoavelmente proporcional dos demais departamentos para a instituição como um todo. Além disso, pela visualização das principais palavras classificadas através de *machine learning*, conseguiu-se inferir que a instituição vem publicando temas voltados para Física Quântica, Nutrição, Religiosidade, Medicina e Engenharia como um todo, o que deixa portas abertas para futuros estudos, posto que algumas das palavras exibidas apresentam sentidos específicos, enquanto outras mais amplos ou genéricos, o que permite resgar informações valiosas para gestão e governança de conhecimento.

Palavras-chave: *Tex Mining*, Ciência de Dados, Governança do Conhecimento

ABSTRACT

It is commonly spread nowadays that today's age is data age. Although it is important to understand the differences between data, information, knowledge, and wisdom. Throughout the years it gets more and more evident that the challenge in knowledge development is not the lack of data but the opposite: its huge amount - so as it is called: big data. Considering that the formal knowledge production occurs - in its majority - inside Universities, today it is clearly understood that, in order to it happens efficiently, a coordination of systemic factors is necessary, and for those it should be approached by a Knowledge Governance point of view. This paper, thru exploratory analyses and application of Data Science tools, intends to highlight information once known only in a tacit way for UFJF. It also aims to offer a direction in which the institution's research are heading for and it is capable of providing quantitative and qualitative analyses over the University's actual scenery. In addition, by applying computer and statistical tools it indicates the areas in which UFJF is considered a reference, the main subjects that were mostly published during the last twenty years, which areas are responsible for the largest numbers of publications, and it shows – in a wordcloud format – those which were classified as most relevant by machine learning algorithms. The methodology was divided into data extraction thru Web of Science, then manipulation using Excel and Python. In excel it happened the cleaning and exploratory analysis, while in Python the manipulation was more complex and it involved data mining, text mining, machine learning and wordclouds. After these process' conclusion, it was possible to identify that Physics, Chemistry and Electrical Engineering were the areas of UFJF that contributed the most when it comes to publication numbers. It was also possible to perceive that there is a proportional distribution from all the institution's departments as a whole. Beyond that, by the visualization of the words exposed in the wordclouds, it could be concluded that the university has been publishing subjects related to Quantum Physics, Nutrition, Religion, Medical and Engineering. That leaves future studies open opportunities because while some of the words present more generic meanings, others present more specifics, which allow value information for Knowledge Management and Governance applications.

Keywords: Text Mining, Data Science, Knowledge Governance

LISTA DE FIGURAS

Figura 1- Pirâmide do conhecimento 22
Figura 2 - Modelo esquemático de governança e gestão do conhecimento sob a perspectiva institucional 26
Figura 3- Mecanismos de GovC adaptados ao âmbito institucional 37
Figura 4- Diagrama de Venn aplicado à Ciência de Dados 46
Figura 5- Etapas da Descoberta de Conhecimento em Bancos de Dados 50
Figura 6- Utilização de linguagens de programação em Ciência de Dados 54
Figura 7- Recomendação de linguagens para ciência de dados 55
Figura 8- Hierarquia horizontal de Data Mining 61
Figura 9- Curva logística com relação bem definida 64
Figura 10 - Representação gráfica de Support Vector Machine 66
Figura 11 - Dez setores em maior número de publicações..... 70
Figura 12 - Nuvem de palavras regressão logística utilizando countvectorizer..... 72
Figura 13 - Nuvem de palavras regressão logística TF - IDF 75
Figura 14 - Nuvem de palavras baseada em Naive Bayes..... 75
Figura 15 - Nuvem de palavras baseada em SVM..... 77
Figura 16 - Principais áreas e seu percentual de participação..... 78

LISTA DE QUADROS

Quadro 1- Tipos de aprendizagem e suas fases..... 59
Quadro 2 - Aplicações mais comuns de algoritmos aos tipos de aprendizagem.... 61
Quadro 3- Glossário das colunas presentes na base de dados da WoS 67

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

IES – Instituição de Ensino Superior

P&D – Pesquisa e Desenvolvimento

GovC – Governança de Conhecimento

CCDG – Centro Canadense para Desenvolvimento de Gestão

TI – Tecnologia da Informação

TIC – Tecnologia da Informação e Comunicação

ERP – *Enterprise Resource Planning*

SCM – *Supply Chain Management*

CRM – *Customer Relation Management*

IoT – *Internet of Things*

BI – *Business Intelligence*

IBPAD – Instituto Brasileiro de Pesquisa e Análise de Dados

DCBD – Descoberta de Conhecimento em Bancos de Dados

PLN – Processamento de Linguagem Natural

SVM – *Support Vector Machine*

SUMÁRIO

1. INTRODUÇÃO 22

1.1 JUSTIFICATIVA 27

1.2 ESCOPO DO TRABALHO 28

1.3 OBJETIVOS 29

1.4 DEFINIÇÃO DA METODOLOGIA 29

1.5 ESTRUTURA DO TRABALHO 31

2. REFERENCIAL TEÓRICO 33

2.1 GOVERNANÇA E GESTÃO DO CONHECIMENTO E CAPACITAÇÕES DINÂMICAS SOB A ÓTICA DAS IES 34

2.2 CIÊNCIA DE DADOS, DATA E TEXT MINING 42

2.2.1 *DATA MINING* 47

2.2.1.1 COMPUTAÇÃO 54

2.2.1.2 MATEMÁTICA & ESTATÍSTICA 58

2.2.1.2 CONTEXTO DE APLICAÇÃO 66

3. DESENVOLVIMENTO 68

3.1 MANIPULAÇÃO EM EXCEL 68

3.2 MANIPULAÇÃO EM PYTHON 70

4. ANÁLISE DOS RESULTADOS 81

5. CONCLUSÕES 83

REFERÊNCIAS 84

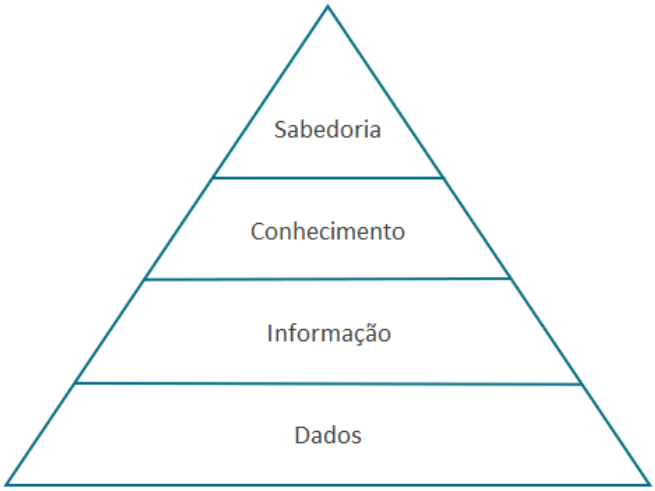
1. INTRODUÇÃO

A noção da importância da informação para a realidade das empresas se tornou senso comum, uma vez que a ideia exposta por Drucker (2007) na metade do século passado de que "o que não pode ser medido, não pode ser controlado" já se encontra enraizada tanto na cultura ocidental quanto na oriental. A facilidade, presente hoje, de captação e acesso a dados permite diversos benefícios; entretanto, de acordo com Fonseca (2019), os desmedidos volumes em que esses se apresentam acabam por gerar novos desafios no seu uso útil.

Em seu trabalho, Jennex and Bartczak (2013) distinguem dados, informação, conhecimento e sabedoria. Eles definem dados como uma coleção de fatos básicos e discretos, enquanto o conceito de informação é descrito como a relação que dados têm entre si, em um determinado contexto, permitindo o desenvolvimento de uma história fundamentada. Já conhecimento é dito como a informação culturalmente entendida que fornece explicações para o motivo e modo como um assunto pode ser aprendido. Por sua vez, sabedoria é entendida como o conhecimento posicionado em uma rede de maneira estruturada, podendo ser aplicado a situações intuitivas ou não.

Como forma de indicar a maneira hierárquica em que estão agrupados, Ackoff (1989) utilizou o formato geométrico de uma pirâmide, onde os dados compõem a base e as demais fatias vão se tornando menores, passando pela informação, conhecimento, até chegar no topo representado pela sabedoria (Figura 1). Em outras abordagens, sobretudo naquelas pertinentes ao escopo do presente trabalho, sabedoria também pode ser compreendida como aprendizado.

Figura 1- Pirâmide do conhecimento



Fonte: adaptado de Akoff (1998)

A engenharia de produção - por ser capaz de atuar no controle, otimização e predição nos mais diversos sistemas - está intimamente relacionada com a coleta e utilização dos dados. Essa relação entre engenheiros e a informação vem mudando e, de acordo com Sharda, Delen e Turban (2019), foi evoluindo e agregando ferramentas ao longo das décadas. Os mesmos autores utilizaram um recorte temporal de início nos anos 70 até o presente momento, onde foi possível caracterizar cada uma das décadas com as práticas de utilização de dados, como segue:

a) Sistemas de apoio à decisão (década de 70): nesse período os sistemas de informação eram majoritariamente compostos por relatórios de rotina de diferentes periodicidades (diários, semanais, mensais etc.). Estes já eram capazes de gerar um elevado volume de dados, o que demandou a criação de ferramentas computacionais e matemáticas (como pesquisa operacional), que fossem além dos recursos intelectuais apenas do ser humano, originando os sistemas de apoio à decisão. Tais sistemas exigiram maior inserção de profissionais aptos a enxergar padrões, processos e opções de otimizações, - engenheiros de produção ou industriais são um exemplo destes profissionais;
b) Sistemas de informação (década de 80): essa década, no contexto de dados, é marcada pela integração dos diferentes sistemas de informação (contábil, marketing, produção, logística, finanças) em um único, ocasionando a aparição dos primeiros ERP (*Enterprise Resource Planning*). Sistemas capazes de apresentar relatórios com versatilidade, porém com manutenção de integridade dos dados, posto que estão todos interligados em um único conjunto computacional.
c) Inteligência de negócios (década de 90 e anos 2000): nesse período surgiu a necessidade de relatórios projetados exclusivamente para os executivos e tomadores de decisão. Nestes estariam contidas as informações primordiais de maneira visual, fácil e ágil, entretanto, com sua confiabilidade garantida. Foi criado um local "intermediário" para armazenamento dessas informações destinadas aos executivos, os chamados locais de armazenamentos de dados fixos (*Data Warehouse*), além da popularização da utilização de *dashboards* e *scorecard*, com o intuito de facilitar a transmissão da mensagem por trás dos dados. Com o advento da globalização, a competitividade do mercado também se tornou global, o que

demandou agilidade na extração, visualização e interpretação dos dados. Para lidar com o atraso de informação, já nos anos 2000, foram desenvolvidos os sistemas de armazenamento de dados em tempo real;

d) *Big Data* (a partir de 2010): o cenário atual apresenta quantidade colossal de dados produzidos, coletados e armazenados; sendo estes em tempo real e com armazenamento em nuvem. A partir daí, fez-se necessária nova quebra de paradigma, culminando na aplicação de técnicas de computação e estatística - como *Data* e *Text Mining*, *MapReduce*, *NoSQL* – para extração de informação válida a se transformar em conhecimento.

Fica perceptível que percorrer o caminho da base ao topo da pirâmide de Askoff (1998) – ou seja, a partir dos "dados" alcançar a "sabedoria"-, é atividade que demanda planejamento por parte dos produtores e usuários de conhecimento. Adicionalmente, posto que o cenário atual demanda também aplicações de ferramentas voltadas para a manipulação eficiente de tais dados, Fonseca (2019) aborda que, em um período razoavelmente recente, a literatura começou a considerar que a transformação dos ativos de conhecimento[1] em inovação (uso "útil") não deve acontecer apenas de forma espontânea, aleatória ou sem coordenação. Isto induz o viés estratégico da sua produção, e mais propriamente, à aplicação de esforços direcionados e intencionais.

Nigro (2016) complementa que a produção de conhecimento formal acontece majoritariamente nas Instituições de Ensino Superior (IES) e de pesquisa. E, por tal razão, Fonseca (2019) também defende que a produção eficiente e eficaz de conhecimento pelas mesmas, demanda a atuação coordenada de fatores sistêmicos. Em outras palavras, de uma ação articulada com as necessidades do entorno, sendo este composto por outras instituições (de pesquisa, apoio, intermediação), governos, empresas e sociedade como um todo. Tal pano de fundo é nomeado por Governança do Conhecimento sistêmica.

A revisão bibliográfica apresentada neste estudo parte desta perspectiva macro e externa para justificar a necessidade da Gestão do Conhecimento em IES. Entende-se como tal a coordenação dos processos e recursos intangíveis intrínsecos à instituição, respaldados pelas suas interações. Rubenstein-Montano (2001) aponta que ambas abordagens (governança e gestão do conhecimento) não correspondem apenas às iniciativas pontuais, e dependem de melhoria incremental contínua. Adicionalmente, as demandas por

[1] Entende-se como os componentes do portfólio científico e tecnológico de uma instituição ou organização.

conhecimento do meio e a dinâmica do seu fluxo mudam constantemente. Para Freire *et al* (2017), isso torna necessário que a Governança do Conhecimento apresente como pilares transparência, equidade, prestação de contas e responsabilidade.

A tratativa da Governança do Conhecimento envolvendo os arranjos institucionais ainda é escassa na literatura, no entanto as aplicações no âmbito das organizações serão brevemente apresentadas, posto que algumas práticas, após sofrerem adaptações, podem ser aderentes à perspectiva sistêmica. Fresneda (2008) indica que o fato das instituições atenderem à sociedade como um todo dificulta a clareza na definição de um cliente final, como ocorre nas organizações privadas. Logo, o fluxo de conhecimento também se apresenta diferente dentro e através das fronteiras de uma IES. As recentes políticas públicas, - como o Marco de Ciência, Tecnologia e Inovação de 2018 e o projeto FUTURE-SE -, caminham de modo a estreitar as relações entre instituições de pesquisa, empresas e sociedade, embora com diretrizes divergentes e, em alguns aspectos, controversas. Processo esse que auxilia no direcionamento desse fluxo, ajudando a transformá-lo em conhecimento aplicável. Por sua vez, o conhecimento de seus pontos fortes, por parte das IES, é peça fundamental no processo de coordenação de trânsito de conhecimento como um todo.

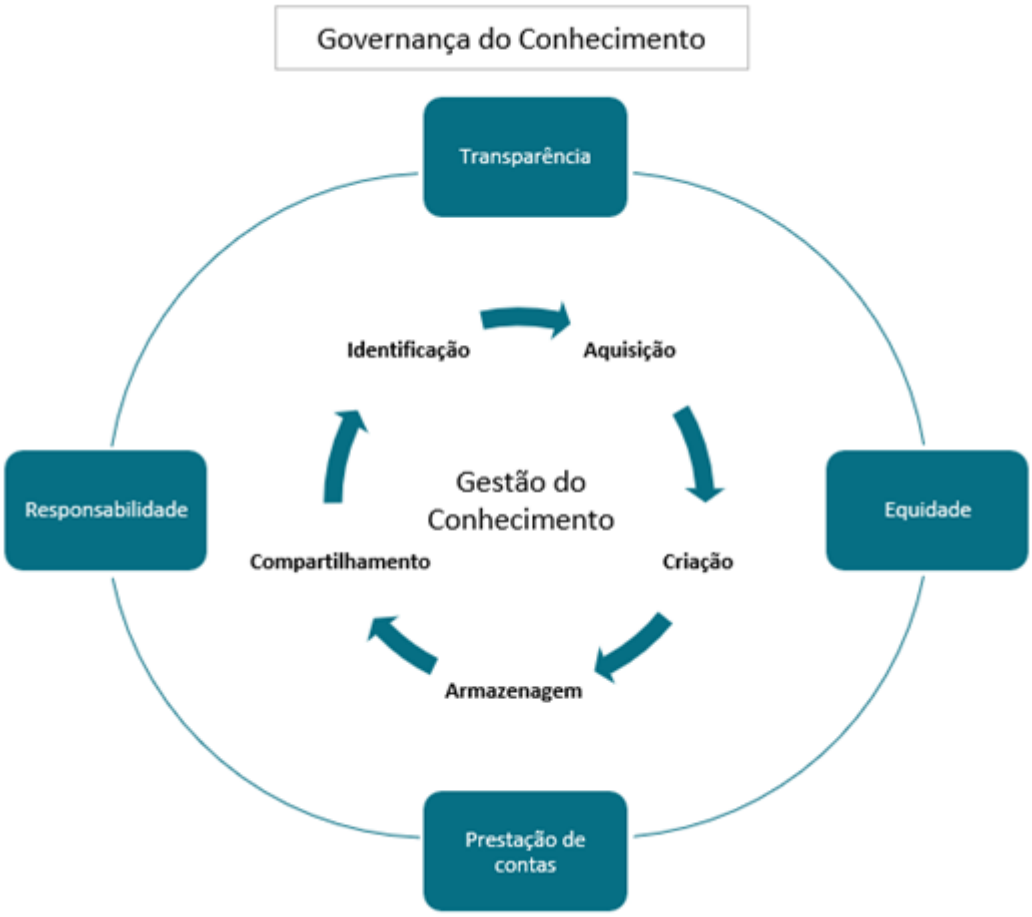
Como já posto, as demandas por conhecimento vindas da sociedade não são estanques. Pelo contrário, mudam de forma cada vez mais acelerada. Alinhada a compreensão de governança supracitada, pode-se empregar a de Capacitações Dinâmicas, defendida por Teece (2007), e também até então mais comumente direcionada às organizações. Do mesmo modo que ocorre nas corporações, as comunidades acadêmicas podem se beneficiar das informações oriundas do meio para orquestrar suas competências em acordo com as necessidades atuais e futuras. Portanto, é importante ressaltar que as capacitações de uma IES também devem ser dinâmicas. No entanto, para que isso ocorra, é mandatório que repensem a maneira como lidam com o armazenamento, identificação e processamento desse grande volume de dados atualmente, começando pelos internos.

Para Leite e Costa (2007), a Gestão do Conhecimento em instituições abrange a coordenação de processos e recursos inerentes à instituição em si. Uma peculiaridade das IES apontada pelos autores corresponde à tratativa da Gestão do Conhecimento dividida em dois planos coexistentes. O primeiro remete ao plano horizontal – correspondendo aos pesquisadores e seus departamentos (podendo estes terem, ou não, interesses compartilhados) – e formam as comunidades acadêmicas. O segundo é definido como vertical e está relacionado ao conhecimento compartilhado de uma disciplina ou tópico de estudo em

específico, apresentando caráter disciplinar – constituindo as comunidades científicas e/ou epistêmicas. É natural que ambos os planos se impactem mutuamente, bem como impactem os fluxos de conhecimento internos e externos, de forma multidirecional.

Freire *et al* (2017) oferecem uma síntese à gestão do conhecimento aplicada ao âmbito institucional, elencando cinco elementos primordiais, a saber: identificação, aquisição, criação, armazenagem e compartilhamento. Vale destacar que é nos dois primeiros elementos que o presente trabalho se debruça. A Figura 2 busca resumir de forma esquemática os pilares de Gestão e Governança do Conhecimento e seus relacionamentos nele empregados.

Figura 2 - Modelo esquemático de governança e gestão do conhecimento sob a perspectiva institucional



Fonte: adaptado de Freire, Dandolini e Souza (2017)

A Ciência de Dados e, neste caso, o levantamento de informações acerca do perfil de publicações por meio da exploração de uma base existente, porém extensa, possibilita extrair informações valiosas sobre o portfólio científico de uma instituição de ensino e

pesquisa, além de fornecer o insumo para estudos de natureza prospectiva. Dalton (2019) sugere que o conceito de Ciência de Dados se insere neste contexto atuando de maneira a fazer o intermédio entre o *Big Data*[2] e os tomadores de decisão destas instituições.

Mendes (2018) infere que o trabalho realizado por um cientista de dados, por envolver ferramentas pautadas em computação, matemática e estatística, aplicadas em um contexto específico, é atividade útil em qualquer organização ou instituição. A capacidade de extração de valor interpretativo em meio a um alto volume de dados possibilita aos gestores nas IES decidir com maior confiança e menor risco, já que estes agora possuem maior embasamento para formulação de suas escolhas.

Com a engenharia de produção, em especial seu campo que trata de tecnologia da informação, - exposta por meio da ciência de dados - atuando no âmbito interno como ferramenta de suporte à gestão do conhecimento e extrapolando ao âmbito externo como auxílio à governança, tem-se aí, por parte da Universidade Federal de Juiz de Fora (UFJF), possibilidade de incremento e também de contribuição mais efetiva na geração de riqueza e conhecimento para a própria instituição e para aquelas localizadas no entorno da região em que se insere.

1.1 JUSTIFICATIVA

O autor se mostra interessado pelo estudo da área de *Data Science* – em especial na utilização da linguagem de programação interpretada *Python* – onde visualiza oportunidade de graduação com o domínio de ferramenta que oferece diferencial e destaque no atual mercado de trabalho. De acordo com Dalton (2019), o desenvolvimento de profissionais que têm conhecimento em linguagens como *Python* e R cresce, em média, 5% ao ano, enquanto a demanda do mercado por esses profissionais cresce à taxa de 21%.

É sabido entre a comunidade de cientistas de dados, que a melhor maneira de aprendizagem de uma linguagem de programação é através de cursos e estudos, porém esta é primordialmente efetiva durante sua aplicação em algum projeto. A opção pelo tema se deu, então, através da união do desejo de vinculação de aplicação de ferramentas de ciência de dados à um projeto atraente para a UFJF. O interesse pelo levantamento de informações sobre o portfólio científico na instituição nasce justamente da urgência aparente ao se pensar

[2] Área de estudo que busca maneiras inovadoras de analisar e extrair informações válidas em meio a bases de dados que são grandes ou complexas demais para serem tratadas por meio de ferramentas tradicionais de processamento de dados.

que esta possui envolvimento, nos últimos 20 anos, em mais de 7500 artigos na *Web of Science (WoS)*, porém não possui ainda um *dashboard* de publicações científicas ou redes de pesquisa bem definido. Criou-se, aí, relação de protocooperação entre o autor, orientadora e UFJF.

O presente trabalho, através de análise exploratória de dados e *Text Mining*, evidencia informações que, de forma geral, existiam de maneira tácita para pesquisadores, professores e gestores da UFJF. Ao tornar mais evidente a vocação de pesquisa da instituição de ensino, este estudo aponta o caminho para o qual as pesquisas da instituição estão indo, além de indicar uma relação não trivial sobre as áreas de pesquisa que nela se destacam. Os resultados do estudo são capazes de fornecer análises qualitativas e quantitativas importantes, para que os tomadores de decisão da UFJF possam ter maior conhecimento do cenário atual – e até de certa forma futuro – sendo assim capazes de traçar estratégias mais assertivas no que compete à IES em questão.

A extração de informação e conhecimento de elevados volumes de dados é prática comum às IES, entretanto, a aplicação de pesquisa computacional, estatística e o processo de análise de redes - a partir de publicações e citações da *WoS* com participação de pesquisadores da UFJF - é atividade até então pouco explorada e que apresenta prognóstico positivo para esta instituição. Oferece-se assim, resultados palpáveis que permitem visualização do panorama das pesquisas da UFJF e de suas comunidades científicas, facilitando a apresentação da instituição para a sociedade e potenciais parceiros, além de indicar as áreas - dentro da IES – que estão mais intimamente relacionadas ou quais são merecedoras de maior atenção e incremento de vinculação.

1.2 ESCOPO DO TRABALHO

Este trabalho, exclusivamente por meio de análise exploratória de artigos da *WoS* e do emprego de ferramentas estatísticas e computacionais de *Text Mining*, ambiciona responder as questões: Em quais áreas a UFJF pode ser considerada referência, com base nas suas publicações científicas? Sobre quais temas a instituição publica mais? Quais as palavras e linhas de estudo chama mais atenção nas publicações dos últimos anos?

Posto que o cenário científico é dinâmico, o recorte temporal utilizado na extração dos dados refere-se às publicações dos últimos vinte anos, o recorte físico engloba apenas a UFJF, optou-se pela seleção de artigos científicos vinculados à *WoS* e inclusão apenas de

trabalhos cuja participação dos pesquisadores da UFJF apresentasse relevância, ou seja, quando figuram entre os seis principais autores do artigo. Estes foram os filtros aplicados com o intuito de geração da base de dados utilizada para extração das respostas das perguntas previamente formuladas, filtros estes que permitiram a extração de 7500 artigos dos mais variados temas, áreas e autores. O primeiro passo para legitimação de análise de dados é a confiabilidade de sua base, logo a fonte, bem como a determinação temporal e os filtros aplicados atendem os requisitos impostos para geração de resultados válidos.

O presente trabalho, por conta de limitações de tempo e de ausência de outras fontes de dados – como os currículos Lattes, por exemplo - limita-se a responder as indicações supracitadas, adentrando, ao máximo, em questões como *Text Mining* e análises embasadas em fundamentos de *Machine Learning* através de algoritmos de classificação.

1.3 OBJETIVOS

O enfoque macro do presente trabalho foi em levantar e analisar informações sobre o portfólio científico da UFJF com o emprego de ferramentas de *Text Mining*, de maneira a responder aos questionamentos evidenciados no escopo. Entretanto, a identificação desse portfólio incluiu consecução de etapas prévias que podem ser entendidas como micro etapas ou objetivos menores. Alguns desses objetivos merecem destaque, como se evidencia:

- Extrair e limpar a base de dados de modo que fosse possível identificar as variáveis-chave do estudo por meio de análise exploratória;
- Selecionar, empregar e avaliar o uso de ferramentas de *Data/ Text Mining* e *Machine Learning* para a finalidade proposta;
- Desenvolver um script em Python composto por todas as ferramentas adotadas que pudesse ser usado com novas bases de dados no futuro.

1.4 DEFINIÇÃO DA METODOLOGIA

Ambicionando o atendimento dos objetivos já descritos e com intenção de elaboração de respostas às perguntas iniciais, o trabalho está pautado em aplicação de pesquisa que envolvem as áreas de computação, estatística e matemática.

Para tal, etapas sequenciais são realizadas, onde cada uma se utiliza de técnicas e ferramentas particulares. De maneira didática, fica evidenciada cada uma das grandes etapas e suas técnicas principais relacionadas como se segue:

1º) Contextualização do processo de produção de conhecimento em uma IES: o enfoque se dá através identificação da maneira como o conhecimento é gerado e administrado em uma IES. Para esse entendimento foi necessário estudo sobre temas que concernem à Governança de Conhecimento, Gestão de Conhecimento e capacitações dinâmicas em IES.

2º) Identificação de base de dados confiável e extração: utilizou-se de experiência prévia para a escolha e manuseio da base. A extração ocorreu através do *website* da *WoS* já de maneira a apresentar o formato adequado (estruturado) para manipulação futura em Excel e *Python*. O acesso se deu através do portal periódicos CAPES e sua plataforma CAFe. Dentro da plataforma incluiu-se a UFJF como um filtro para instituição de ensino e restringiu-se como resultados apenas os artigos. Após a exibição dos resultados, a extração da base ocorreu clicando no botão "exportar para outros formatos de arquivo" e selecionando-se 500 registros por vez (máximo ofertado pela plataforma), opção por registro completo e formato separado por tabulações (WIN). O processo precisou ser repetido até o esgotamento dos artigos. Os arquivos foram baixados em formato txt, porém puderam ser exibidos em uma planilha do Excel, logo foram compilados seguidamente um abaixo do outro formando um arquivo único (incluindo um cabeçalho) e salvos em formato xlms ou CSV.

3º) Preparação e limpeza dos dados: foram feitas através de ferramentas próprias ao Excel como renomeação de colunas, divisão de colunas, agrupamento de informações, aplicação de filtros, preenchimento de dados faltantes, entre outros. A limpeza também incluiu remoção de colunas que continham informações desinteressantes ao trabalho, bem como adição de novas contendo dados não evidentes. Ademais, foi feito o preenchimento de dados faltantes com a palavra "Vazia" (o que contribuiu na manipulação futura em *Python*)

4º) Análise exploratória: essa análise inicial ocorreu em Excel, onde o autor foi capaz de extrair gráficos e relações evidentes e - não tão evidentes – que

permitiram a visualização de possibilidades a se explorar. Nesta etapa, posto que não existia clareza ainda com o que esperar dos dados, foram realizadas diversas associações e gerados variados gráficos com o intuito de formulação de *insights* para aplicações futuras. Vale destacar, que para as etapas três e quatro o autor pôde optar por utilizar Python, entretanto escolheu Excel por conta de sua praticidade e familiaridade maior com o *software*.

5º) Text mining: já se utilizando de *Python*, o autor aplicou ferramentas de bibliotecas vinculadas ao software para realizar nova limpeza (porém de dados textuais) e encontrar informações referentes aos textos, como a incidência de palavras-chave, identificação de expressões mais utilizadas, etc. Tal etapa também é conhecida na literatura como pré-processamento.

6º) Análise preditiva: novamente através da linguagem *Python*, foram aplicados algoritmos de *machine learning* (também encontrados nas suas bibliotecas) para fins de classificação e associação (a exemplo da vinculação dos termos mais empregados ao número de citações e visualizações dos artigos).

7º) Elaboração do portfólio: após reunidas as informações, foi possível a elaboração de nuvens de palavras e gráficos, de modo a tornar mais evidente o cenário atual de pesquisa da UFJF, bem como indicativo de possibilidade de cenário futuro.

1.5 ESTRUTURA DO TRABALHO

Este trabalho está dividido em cinco capítulos, onde o capítulo inicial corresponde à introdução do tema, e é apresentada a relação não trivial de geração e gestão de conhecimento interno e externo em uma IES com ciência de dados. O capítulo também evidencia a motivação do autor pela sua realização, além de exaltar os objetivos pretendidos, as etapas que foram seguidas e as ferramentas das quais se utilizou.

O capítulo seguinte corresponde à revisão bibliográfica que se inicia explanando conceitos como governança e gestão de conhecimento, além de capacitações dinâmicas em IES. Em seguida, explora a relação desses temas com ciência de dados, além de trazer histórico da evolução deste último, especialmente suas questões mais atuais que envolvem elevados volumes de dados – o chamado *Big Data*. Ao se referir a cientistas de dados,

recapitula os três grandes campos (ou habilidades primordiais), onde traz conceitos como *data* e *text mining*, além das ferramentas estatísticas e computacionais pertinentes – em especial a linguagem Python e algumas de suas bibliotecas.

O terceiro capítulo se refere ao desenvolvimento, ou seja, é exibido o passo a passo do desenvolvimento do trabalho, bem como informado o como e o porquê de utilização de cada ferramenta estatística e computacional. Enquanto o quarto exibe os resultados planejados, que pode ser entendido como o resultado da conclusão da análise exploratória das bases, as redes criadas a partir da informação contida nas bases tratadas e análise preditiva após aplicação de algoritmos de *machine learning*.

Finalizando, compete ao último capítulo expor as conclusões elaboradas pelo autor após finalização do estudo, além de deixar direcionamento das oportunidades futuras de trabalho.

2. REFERENCIAL TEÓRICO

Na visão de Souza e de Moraes (2012) a palavra conhecimento remete à congregação de informações que foram adquiridas por experiências em aprendizagem guiada ou vivência. Essas informações, após compreendidas, podem ser difundidas de maneira informal ou formal, onde a última comumente está relacionada às instituições de ensino. Lozada *et al* (2016) entendem, então, que o conhecimento formal corresponde ao método estruturado que segue um planejamento e apresenta uma intenção ou objetivo final com sua transmissão.

Conforme Fonseca (2019) evidencia em seu trabalho, o desenvolvimento efetivo de conhecimento dentro de uma instituição carece de esforços organizados e planejamento. Já Leite e Costa (2017), ao avaliar a produção de conhecimento por parte de uma IES, destacam que essa coordenação e planejamento de atividades acontece tanto no âmbito interno – envolvendo pesquisadores e recursos dos departamentos internos das IES, bem como no âmbito externo – ao evidenciar a relação que uma instituição apresenta com outras em seu entorno ou até regiões distantes. Relação essa possível e intensa através do fenômeno da globalização.

Além de globalizado, o mundo também se encontra imerso em tecnologia. Vydiswaran (2017) informa que o volume de geração de dados cresce exponencialmente, sendo esperada a produção de 40 *Zettabytes* (40 bilhões de *Terabytes*) por dia no ano de 2020. Dados estes, conforme Askoff (1998) indica, que são justamente os constituintes da base de construção de conhecimento.

Impossível se faz, então, dissociar-se uma IES com a utilização de dados. Vydiswaran (2017) explana que 80% dos dados se apresentam de maneira não estruturada, especialmente os ricos em textos. Ao se pensar na relação dos dados textuais com as IES, destaca-se a produção de mais de 40 milhões de artigos científicos, sendo destes 8500 contendo participação de pesquisadores da UFJF nos últimos vinte anos, apenas considerando os disponibilizados na *Web of Science*.

Pelo fato de a sociedade estar imersa numa era de altíssimos volumes de dados – o chamado *Big Data* -, Dalton (2019) destaca a importância da ciência de dados e suas ferramentas para extração de informação útil em meio a esse vasto universo. O que ocorre de maneira análoga por parte de uma IES com foco em efetividade de produção de conhecimento.

A partir desta necessária articulação, os conceitos de governança e gestão do conhecimento, capacitações dinâmicas e ciência de dados são detalhados a seguir. Após esse detalhamento, o último conceito será aprofundado com a apresentação das ferramentas de *Data Mining* e *Text Mining*, da linguagem de programação interpretada Python e as particularidades de algumas de suas bibliotecas e métodos estatísticos pertinentes ao estudo.

2.1 GOVERNANÇA E GESTÃO DO CONHECIMENTO E CAPACITAÇÕES DINÂMICAS SOB A ÓTICA DAS IES

Nonaka e Takeuchi (2008) contribuem com a discussão sobre conhecimento formal e informal ao conceituarem conhecimento tácito e explícito e a relação dinâmica entre ambos. O primeiro pode ser entendido como aquele adquirido pelo indivíduo ao longo de suas experiências de vida e vivências, enquanto o segundo é aquele que pode ser registrado de forma estruturada, ou seja, seguindo regras e sendo transmitido através de uma espécie de codificação. No que tange à relação destes, os mesmos autores enfatizam que ambos não podem ser dissociados, posto que interagem de maneira dinâmica e cíclica.

Conforme análise de Leite e Costa (2007), a transmissão de informação e posterior geração de conhecimento científico, envolve uma série de esforços para se desenvolver processos dinâmicos capazes de coordenar a relação entre conhecimento explícito e tácito presentes dentro e fora da academia. A comunicação entre os acadêmicos em si e entre os acadêmicos e a sociedade é o fator base de transmissão de conhecimento científico. Em se tratando de comunicação, os autores também evidenciam que, apesar da comunicação formal ser a mais apropriada ao campo de produção científica, a informal é a mais presente nesse cenário, sendo essa a principal contribuidora para manutenção de fluxo de informação. Tal fato se dá, pois é natural concluir-se que um pesquisador direciona seu foco e objetivos de pesquisa de acordo com os interesses que foram construídos ao longo de sua vida: de maneira explícita - dentro do universo acadêmico - e de maneira tácita - dentro das IES e igualmente no contato com seus grupos sociais não acadêmicos.

O desenvolvimento do conhecimento organizacional deriva da gestão de práticas de estímulo à criatividade e interação dos colaboradores, da capacidade de adaptação da organização ao ambiente, aos investimentos em Pesquisa e Desenvolvimento (P&D) e na produtividade alcançada. Logo, Foss (2007) traz a ideia da *Knowledge Governance Approach* (KGA), onde entende-se a Governança de Conhecimento (GovC) aplicada às organizações como conceito de aplicação interdisciplinar de diferentes áreas da gestão, sendo

elas a gestão do conhecimento, gestão estratégica, gestão de recursos humanos e entendimento de estudos organizacionais com o intuito de otimização da produção e aquisição de conhecimento (incluindo os oriundos das instituições de ensino e pesquisa).

Com o interesse de se pensar numa abordagem que envolva o âmbito público e privado simultaneamente, ou seja, sistêmica, Burlamaqui (2011) reformulou a conceituação do termo GovC. O mesmo autor enxerga a governança como opção para reduzir as divergências entre interesses públicos e privados e até mesmo os interesses públicos em diferentes instituições. Isso ocorre, pois a GovC promove cooperação entre as partes - podendo essa ser entre duas ou mais IES, entre uma IES e outra instituição pública, entre uma IES e uma organização e até mesmo entre uma IES e a sociedade. O termo, sob este enfoque, também fornece diretrizes para coordenação das ações que ambicionem o bem coletivo público concomitantemente com o bem empresarial.

Além da conceituação da GovC no meio organizacional, faz-se necessário entender a maneira como a governança de conhecimento atua sob a ótica institucional. Segundo Freire *et al* (2017) o seu modelo de aplicação remete a outros modelos de governança preexistentes (público, global e corporativo). O enfoque deste, porém, está direcionado ao equilíbrio entre dependência e poder; ferramentas de manutenção da capacidade de aquisição de conhecimento; fomento à comunicação entre pesquisadores internos e atores externos às instituições; e, estratégias de captação e agregação do máximo das capacidades técnicas pelos grupos de pesquisa.

O modelo de GovC institucional, para ser possível, demanda adaptação de princípios e mecanismos do modelo organizacional. A partir de compilação da literatura da área, Freire *et al* (2017), didaticamente, dividiram a GovC em quatro categorias, cada qual fundamentada em seus mecanismos. Observa-se, a seguir, o seu detalhamento, e posteriormente sua estrutura hierárquica (ilustrada na Figura 3).

a) Transparência: para Grandori (2009) *apud* Freire *et al* (2017) é necessária a criação de redes e parcerias inter e intra institucionais com o intuito de estimular a comunicação, agilizar o fluxo de informação, aproximar distâncias cognitivas e assim facilitar a descoberta de interesses em comum, especialmente no âmbito regional. Vale destacar que o funcionamento eficiente das parcerias depende do bom relacionamento entre instituições, pautado em uma cultura de transparência. Ademais, na visão de Foss e Klein (2008), é importante que os responsáveis pela gestão de pessoas formulem

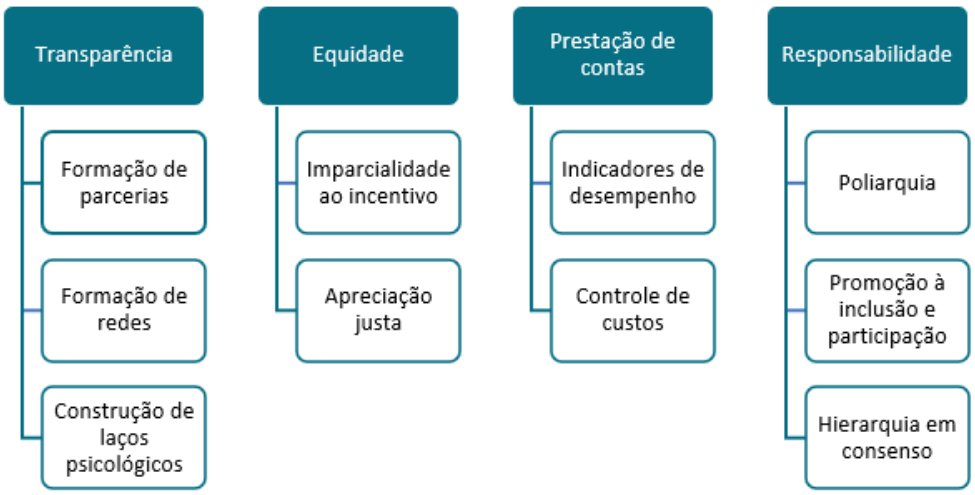
ações que aumentem o desenvolvimento de laços psicológicos e que propiciem confiança ao compartilhamento de informação. Esse mecanismo torna possível a aproximação e maior compreensão entre os pesquisadores, colaboradores e alunos através do crescimento da empatia.

b) Equidade: Grandori (2009) *apud* Freire *et al* (2017) evidencia a relevância do desenvolvimento de ferramentas que estimulem a capacidade de aquisição de conhecimento de maneira justa nas mais diversas áreas de pesquisa, departamentos e instituições. Isso engloba criação de mecanismos políticos que apresentem imparcialidade ao incentivo das diferentes áreas de pesquisa – através de subsídios, bem como promoção de espaço justo para suas divulgações.

c) Prestação de contas: ao abordar o tema com enfoque financeiro, Foss (2006) *apud* Freire *et al* (2017) traz a atenção para a relevância da criação de indicadores de desempenho capazes de relacionar os investimentos aos padrões de entregas esperados a cada instituição. Ao serem utilizados de maneira eficiente, espera-se que sirvam de base para o monitoramento de custos; especialmente aqueles envolvidos com o controle de risco, sendo este inerente ao fluxo de informações.

d) Responsabilidade: Grandori (2009) *apud* Freire *et al* (2017) destaca que instituições eficazes na produção de conhecimento aplicam a ideia de gestão descentralizada (poliarquia), onde as próprias equipes coordenam de forma mais eficiente seus respectivos projetos. Tsuyuki (2006) complementa que essas instituições têm características comuns relacionadas ao sucesso nas práticas de GovC - como a divisão de responsabilidade pautada na adoção de hierarquia consensual e a constante promoção à inclusão e participação de novos pesquisadores e alunos. Ademais, Foss (2006) adiciona a necessidade de inserção de um sistema de liderança em que os tomadores de decisão estejam norteados dentro de uma construção social de sentido, ou seja, façam escolhas que, sabidamente, são de interesse da maioria dos seus colegas e que estejam de acordo com a cultura da instituição, portanto, decisões que estejam fora de vontades individuais.

Figura 3- Mecanismos de GovC adaptados ao âmbito institucional



Fonte: adaptado de Freire, Dandolini e Souza (2017)

Existem tanto barreiras internas, quanto externas inseridas no contexto da governança de conhecimento aplicado às IES. Para Freire (2012) a ausência de linguagem compatível para compartilhamento de conhecimento, bem como o conhecimento tácito de cada indivíduo e grupo formam o que denomina barreiras estruturais. Superá-las, para Alves e Barbosa (2010), significa incentivar os diálogos interdepartamentais, promover aumento da interação entre instituições, organizações e sociedade, em conseguinte legitimar o processo de compartilhamento científico.

Eisenhardt e Martin (2000) entendem que o conhecimento é um recurso fundamental para geração de valor, seja em uma instituição ou organização. Como observado, as dificuldades de identificação deste recurso já estão presentes em um cenário "estático", porém os desafios são ainda maiores ao se pensar numa realidade dinâmica. Essa última é a realidade vivida por todas as instituições, já que todas estão inseridas em um contexto científico de constante inovação, agilidade e transformação. Segundo os mesmos autores, a maneira contínua como os gestores e suas equipes empregam o conhecimento, combinam-o e são capazes de coordenar a geração de novos, se mostra uma opção para o tratamento dos entraves nas IES. Essa aplicação envolve uma ideia empregada no meio organizacional há muitas décadas, atualizada por Teece (2007), mas que podem ser adequadas para o meio institucional.

O conceito de capacitações dinâmicas pode ser entendido pelo desenvolvimento de competências através da aprendizagem social, tornando possível continuamente controlar,

criar, expandir e melhorar os conhecimentos internos e, a partir daí, reconfigurá-los de maneira a desenvolver novas oportunidades (TEECE, 2007). O mesmo autor o divide em três processos que coordenam o dinamismo das capacitações chamados de *sensing*, *sizing* e gerenciamento de ameaças e reconfiguração. Cada uma delas é aqui contextualizada de acordo com a realidade institucional:

a) *Sensing*: pode-se entender como a sensibilidade que a IES possui para captação de novas oportunidades ou sensibilidade para observação de ameaças. Essa é dependente do acesso à base de dados armazenados, que consequentemente correspondem à parte da informação e do conhecimento que possui. Pode ocorrer através pesquisa, por parte da comunidade acadêmica, do cenário corporativo vigente e, assim, entendimento do conhecimento de suas demandas atuais e futuras. Ocorre também de maneira direta, onde não é necessária a dita "sensibilidade", já que a demanda de produção de conhecimento específico é levada à instituição através de contato direto de alguma organização. Ao se pensar nas ameaças, o meio corporativo também direciona seu foco aos concorrentes, porém no meio institucional o olhar está voltado para mudanças nas normas regulatórias e na sociedade como um todo.

b) *Seizing*: representa a captura de oportunidades através de *timing*. Esse pode ser visto como o momento ideal para investimentos ressonantes com a tecnologia de destaque no momento, ou tomadas de decisão alinhadas com as estratégias de inovação vigentes na conjuntura. Para o tal, é primordial conhecimento do cenário corporativo, pesquisas de mercado e interesse pelas necessidades sociais, o que habilita às IES a estarem aptas a desenvolverem e combinarem tecnologias e conhecimentos capazes de agregar valor. Nesse aspecto insere-se também a inovação que impulsiona a habilidade de contornar adversidades e desenvolvimento de pensamentos que fogem ao conhecimento tradicional em uma instituição.

c) Gerenciamento das ameaças e reconfiguração: abrange a capacidade que uma IES possui de reconfigurar suas estruturas e grupos de pesquisas de maneira a explorar o conhecimento pré-adquirido em uma nova perspectiva, adotando novas direções de produção científica e adaptando-se às novas exigências neste cenário dinâmico. Vale ressaltar que a reconfiguração,

quando drástica, pode representar rompimento com a identidade dos campos de pesquisa vigentes considerados “tradicionais”, o que pode ser visto como uma ameaça aos olhos mais conservadores.

O Centro Canadense para Desenvolvimento da Gestão (CCDG, 2001) identificou que as instituições de tamanho considerável, em um geral, apresentam entraves no gerenciamento de conhecimento. Os entraves já seriam desafiadores por si só em um cenário não dinâmico, porém Teece (2007) deixa ainda mais claro o incremento das dificuldades em ambientes de rápida movimentação de informação e caracterizados por dispersão das fontes de produção de conhecimento, ou seja, o típico ambiente de uma IES. Nessas detectou-se uma falha em comum – obscuridade na identificação das necessidades das partes interessadas, dificuldades de comunicação, bem como turvo reconhecimento de suas vocações.

Os pesquisadores Fresneda e Goulart (2007) elencam alguns dos fatores causais dos entraves, ao exemplo do elevado volume de informações, subutilização de competências e diminuta colaboração e trabalhos em grupo dentro das instituições. Fonseca (2019) complementa ao abordar a questão de que o conhecimento é criado também fora da academia, onde o conhecimento tácito obtido por um aluno durante uma aula poderá ser transmitido para sua comunidade ou grupo de convivência, a partir daí esse grupo também pode desenvolver novas ideias sem que as IES jamais tenham consciência.

Como já defendido, uma maneira de lidar eficientemente com tais falhas é a aplicação de ferramentas de Gestão do Conhecimento. A retomada do termo, porém agora utilizando ótica mais profunda e elucidativa, demanda sua conceituação de forma mais robusta, onde Leite e Costa (2007) o descrevem como a coordenação das ações, mecanismos políticos e estratégias que permitem a gestão do fluxo de informação dentro da instituição. Essa coordenação demanda processos que precisam estar alinhados e conectados. Com o objetivo de facilitar a compreensão, os mesmos autores oferecem um modelo que enquadra de maneira pragmática cinco etapas, a saber:

1ª) Identificação: atua de forma a mapear o conhecimento tácito ou explícito já existente na IES, bem como na sua comunidade acadêmica. Através do registro das fontes de geração de dados e produção de conhecimento científico, esta busca identificar as capacitações da instituição, a área de atuação dos pesquisadores e a possível trilha que os conhecimentos futuros irão seguir. O intuito é responder questões que dizem respeito à qual pesquisador pesquisa o

que, em qual aspecto a instituição apresenta vocação e para onde estão rumando suas principais linhas de pesquisa. Conforme adiantado, no cenário atual, dominado pelos altos volumes de dados, essa etapa de aquisição só se faz possível com a utilização e aplicação de ferramentas de ciência de dados, onde ocorre preparação da base de dados e posterior análise exploratória em busca de padrões sequenciais.

2ª) Aquisição: relaciona-se com o processo de obtenção do conhecimento previamente identificado que será armazenado e futuramente transformado no processo de criação. No que tange ao conhecimento explícito, tal aquisição ocorre através das bibliotecas, periódicos, base de dados (relacionais ou não), repositórios da instituição e seus diretórios de patentes. O conhecimento tácito, entretanto, não apresenta possibilidade de ser adquirido de maneira evidente, bem como armazenado.

3ª) Armazenagem: compreende uma estrutura que preconiza organização dos mais diversos conteúdos de maneira a fornecer rápida identificação, recuperação e extração de dados e informação.

4ª) Compartilhamento: para o conhecimento explícito o compartilhamento se dá durante por meio das publicações formais e acesso livre a arquivos; enquanto o tácito acontece de maneira informal através de diálogos e experiências.

5ª) Criação: compreende o desenvolvimento de novo conhecimento científico, ao unir as habilidades e conhecimentos armazenados previamente com os *outputs* dos novos trabalhos dos pesquisadores. Assume-se que a criação em uma IES ocorre através da pesquisa científica. Vale ressaltar que o fator denominado por Takeuchi e Nonaka (1997) como socialização - exposta através de ensino, orientação de qualificações e dissertações – também é fundamental no processo de criação.

Posto que para Nigro (2016) uma das formas de geração de conhecimento formal acontece dentro das Instituições de Ensino Superior (IES), as respostas para as perguntas encontradas após a realização da etapa primeira - Identificação - são fundamentais para o desenrolar do fluxo de fomento à sabedoria. A elaboração dessas respostas, entretanto, é formulada através da exploração de duas diferentes fontes de dados, que Ferraz *et al* (2014) descrevem como bases referenciais e não referenciais. A primeira corresponde às

dissertações, artigos científicos, livros e teses; enquanto a segunda remete aos currículos dos professores/pesquisadores, grupos de pesquisa e as comunidades epistêmicas.

As interações entre os pesquisadores e em conseguinte, a formação de grupos de pesquisa e comunidades epistêmicas derivam, segundo Leite e Costa (2007) do contexto cultural, da história e conhecimento tácito dos pesquisadores envolvidos. Tais comunidades são fundamentais no processo dinâmico de fluxo de informação, além de serem dominantes no âmbito de produção de conhecimento a nível institucional. Entretanto, observa-se que as IES ainda possuem pouca coordenação do que está sendo pesquisado e quais os pesquisadores envolvidos em cada tópico de pesquisa. O gerenciamento dessas redes de pesquisa é atividade a ser desenvolvida internamente nas instituições, podendo ser enquadrado como uma das práticas de gestão de conhecimento.

Ferraz *et al* (2014) enxergam que a utilização das bases referenciais e não referenciais permite mais do que apenas exploração de uma fonte rica de dados, mas a possibilidade de desenho de um panorama da conjuntura de pesquisa de uma IES. Em outras palavras, entende-se que se trata do portfólio científico da instituição a que pertencem. O estudo dessas bases atua de maneira a elucidar quais as tecnologias estão sendo produzidas, como estão distribuídos os grupos científicos e quais os principais focos de pesquisa. Esse mapeamento também permite aclarar a obscuridade presente nas instituições, compondo um primeiro passo onde se evidencia, de maneira pragmática, os tópicos de pesquisa que podem ser considerados suas vocações.

Teece (2007) infere que por conta da característica dinâmica da movimentação dos dados, o que chama “sobe e desce da informação”, as instituições precisam desenvolver mecanismos e procedimentos capazes de manter a gestão sempre informada. No cenário atual, a existência de tais mecanismos só é possível através do uso da tecnologia, em especial uma área específica da Tecnologia da Informação (TI) - a Ciência de Dados.

Ao se pensar na infraestrutura de conhecimento, Fonseca (2019) evidencia um fator de relevância que chamou de “aprender a aprender”. Para tal, faz-se necessário, preliminarmente, observar a existência de lacunas específicas presentes em meio ao universo de informações para, posteriormente, buscar-se seu preenchimento. Esse processo de identificação de lacunas e possibilidade de criação de conexões nasce da análise exploratória de vastas bases de dados.

Batista (2012) propõe, dentre outros itens, a tecnologia como fator crítico de sucesso à gestão do conhecimento. Nesse contexto, pode-se resumir a compreensão de tecnologia às

ferramentas de Ciência de Dados e estas, segundo Dalton (2019), podem ser entendidas como as técnicas que atuam no armazenamento estruturado, tratamento, mecanismos de busca, exploração e disseminação de dados. Assume-se também, a tecnologia como ferramenta de desenvolvimento de redes de aprendizagem, o que permite evidenciação das competências, auxílio na comunicação entre pesquisadores e membros da IES e aumento da interação e trabalho em grupo. Em resumo, a presença de Tecnologia da Informação (TI), em especial a aplicação de computação no contexto de *Data Mining* e *Text Mining*, reduz o *gap* existente entre a base de dados e os tomadores de decisão, além de aumentar a colaboração dos membros de todas as partes interessadas em prol de troca de conhecimento.

2.2 CIÊNCIA DE DADOS, DATA E TEXT MINING

É evidente a importância da produção de conhecimento para uma nação e sociedade. Para tal, Curty e Serafim (2016) caracterizam os dados como insumos fundamentais em dois processos: avanço no desenvolvimento de conhecimento e embasamento para tomada de decisão. Vale destacar, entretanto, que a produção por si só está aquém de atingir sua potencialidade se este insumo não for corretamente armazenado, identificado, transformado e disseminado. Para os mesmos autores, no contexto atual, onde o fluxo de dados é colossal, a presença de TI nas instituições produtoras de conhecimento é mandatória e as instituições dependem diretamente de dados.

As tecnologias de informação e comunicação representam importante papel na divulgação de ciência, criação de conexões entre pesquisadores e velocidade de troca de informação ao redor do mundo. O fato da produção e transmissão de conhecimento ter atingido alcance de escala global culminou na redução de barreiras territoriais anteriormente existentes, porém isso não exclui o fato de as IES apresentarem vocação para produção científica de determinada especialidade para a geração de riqueza em nível local.

De acordo com Stover (2018), no cenário atual, onde a vida é permeada por computadores, satélites, rádio, televisão, internet e outros eletrônicos, as tecnologias da informação e comunicação (TIC) são entendidas pela combinação desses; e sua interação efetiva. Ainda na visão de Stover (2018), as TIC revolucionaram a maneira como trabalhamos, compramos, interagimos e até pensamos. Serviços comerciais, bancários, educacionais, de entretenimento e até mesmo as linhas de montagem são fundamentalmente diferentes do que eram a dez anos atrás por advento da sua aplicação. Por ser um tema de

constante e rápida evolução, acompanhá-lo torna-se tarefa de difícil execução, demandando conhecimento e dedicação. Haberkorn (2018), através de dados da Aberdeen Group (empresa de consultoria americana), estimou que apenas 27,6% da tecnologia prometida é utilizada em contraste à disponibilizada.

Stover (2018), ao elaborar um histórico didático sobre tecnologia da informação, elenca que o desenvolvimento do conceito de banco de dados como parte de TI partiu da necessidade de controle de estoque de materiais e recursos para produção nas unidades de manufatura, dado o crescimento contínuo dos volumes de produção e variedade dos componentes. O sistema evoluiu para coordenar também as ordens de produção, chegando a se tornar um sistema de gerenciamento de banco de dados dos mais diversos recursos da empresa, o denominado *Enterprise Resource Planning* (ERP). Nos últimos anos, foram introduzidas novas funcionalidades ao ERP e incorporados sistemas empresariais para suporte. Esse suporte pode estar direcionado ao fornecedor através da cadeia de suprimentos pelo uso do SCM (*Supply Chain Management*) ou conectado ao consumidor em forma de *e-commerce* e gestão de relacionamento com o cliente por CRM (*Customer Relation Management*).

Essa mudança na intensidade de interação entre as diferentes partes – seja no âmbito organizacional, institucional ou até social como um todo - para Xia *et al* (2012) marca o início de uma nova era nos sistemas de informação e comunicação. Os mesmos autores inferem que o fato de a sociedade estar rodeada de objetos equipados com as mais diferentes funções, porém sempre com conexão à internet e GPS, caracteriza a chamada "*internet of things" (IOT)* ou internet das coisas/objetos. As suas fontes geradoras de dados estão baseadas desde tecnologias complexas como satélites, às mais simples e acessíveis como smartphones, e podem ser interruptas ou contínuas/tempo real. As fontes interruptas ou discretas são as que apresentam maior periodicidade para sua atualização; como aquelas que são atualizadas uma vez ao dia, mensalmente, duas vezes ao ano etc. Na prática, pode-se visualizar como exemplos pesquisas sociais, dados governamentais abertos e as bases de dados extraídas da *WoS* (que inclusive compõem a fonte explorada neste trabalho cujo foco são as IES). Já as fontes contínuas são aquelas atualizadas com tamanha velocidade que são admitidas como "tempo real"; de forma ilustrativa, para essas tem-se: dados climáticos da Terra, redes sociais, sensores, bolsa de valores, entre outros.

Corroborando com os autores anteriores, Atzori *et al* (2010) inferem que o avanço da *IOT* é necessariamente resultado da sinergia entre diferentes áreas de conhecimento como

telecomunicações, informática e a sociedade; ademais, explanam que o aumento considerável da integração dos mais diversos dispositivos de tecnologia e de seres humanos, culmina na criação de redes de comunicação que são constituídas de um enorme volume de dados.

Esse amontoado de dados, justamente por contar com grandes quantidades de informação com potencial de serem transformadas em conhecimento, recebeu denominação própria: os chamados *Big Data*. O novo conceito atraiu atenção, por parte dos profissionais de TIC para sua pesquisa e, em meio a este contexto, os pesquisadores Ward e Baker (2013) procuraram efetuar uma análise sobre esses dados de imenso tamanho e optaram por fundamentá-los em três pilares críticos: o tamanho, a complexidade e as tecnologias. De maneira didática, pode-se visualizar a explicação dos três:

a) Tamanho: corresponde à quantidade de dados, ou seja, a dimensão de seu volume;
b) Complexidade: é definida pela estrutura do conjunto de dados, seu comportamento e a maneira como comutam entre si;
c) Tecnologias: são as ferramentas e técnicas computacionais aplicadas com finalidade de processar a base de dados que apresente elevado tamanho e complexidade.

Fica evidente que para desenvolvimento de tecnologia são necessários recursos e que, ao se adentrar no campo de conhecimento, o principal deles são os dados. A questão de destaque, entretanto, é que os dados se apresentam em volume tão abundante que o planejamento adequado de sua utilização acaba se tornando o componente fundamental do campo atual de estudo de TI. O conceito de ERP, em sua acepção, trata exatamente disso: planejamento de recursos. Como resultado de sua evolução natural, este último passou a compreender - além dos bancos de dados - sistemas de apoio a decisão, como o *Business Intelligence* (BI), *Data Warehouse* e *Data Mining*. Para Kleis *et al* (2012), a TI também pode ser vista como um recurso organizacional que habilita e estrutura a produção de conhecimento, ciência e tecnologia, levando a soluções de inovação fundamentadas em bancos de patentes e bases científicas, por exemplo.

Conforme Bramer (2016) infere, atualmente ocorre um fenômeno que chama de explosão de dados. Os computadores modernos acumulam dados à uma taxa inimaginável, sendo esses oriundos das mais diversas fontes. A título ilustrativo, tem-se uma indicação de que os satélites de observação modernos da Nasa produzem mais dados em um dia do que todos os antigos e desativados já produziram juntos. Estima-se a existência de mais de 650

milhões de sites na web, além de números superiores a 3 bilhões de posts por dia em redes sociais. A questão, na visão do mesmo autor, encontra-se no fato de o mundo estar se tornando rico em dados, mas pobre em conhecimento.

Cury e Serafim (2016) identificaram que o desafio das instituições e organizações não é mais localizar e extrair dados e sim, entender como trabalhá-los de modo a retirar informação útil; ou seja, a dificuldade encontra-se em entender como extrair conhecimento de uma base tão vasta que chega a ser desafiadoramente confusa.

Ambicionando atingir-se o máximo potencial de aplicação de informação extraída das fontes de dados, ocorre a necessidade de utilização de abordagem interdisciplinar que seja capaz de combinar elementos computacionais, matemáticos e estatísticos com enfoque na origem dos dados e com direcionamento de competências específicas de maneira a trazer significado ao final desta análise exploratória. Segundo Loukides (2012), essa abordagem analítica é, justamente, o que compõe a definição de ciência de dados.

Stanton *et al.* (2012) complementa que hoje, após a inserção da interdisciplinaridade de ferramentas computacionais e estatística, ela vai muito além de suas aplicações de coordenação de banco de dados e ordens de produção, permitindo novas perspectivas, posto que oferece análises não só de dados históricos e de otimização, mas também preditivas.

O Fórum Econômico Mundial - *World Economic Forum* (2019) - listou a profissão cientista de dados como uma das mais relevantes até 2020 e, até de maneira cômica, definiu-a como a profissão mais *sexy* do século XXI. Para Grus (2016) pode-se entender que “o cientista de dados é alguém que extrai conhecimento de dados desorganizados”, o que é propriamente o enfoque do presente trabalho ao se pensar em IES.

Para Dalton (2019), apesar do termo *Data Science* (Ciência de Dados) apresentar registros de utilização desde a década de 60, esta é uma ciência que em sua subjetividade bem compreendida é bastante nova. Finzer (2013) indica que a aplicação do conceito ao modo próximo como é entendido hoje data do início deste século, onde de maneira mais ostensiva, aliou-se estatística à programação e computação. O Instituto Brasileiro de Pesquisa e Análise de Dados (IBPAD, 2020) descreve que o conceito atual de Ciência de Dados engloba duas áreas de maneira interdisciplinar: ciência da computação e estatística; entretanto, sua aplicação acontece de forma à servir de suporte aos mais variados campos do conhecimento como economia, medicina, educação, engenharia e outros.

Conway (2010) foi o primeiro a elaborar um pensamento que compreendesse uma terceira competência à ciência de dados: o contexto de aplicação. Anteriormente, o termo

estava relacionado apenas à congruência entre as áreas de matemática/estatística e computação. Na visão do mesmo autor, entende-se que o contexto de aplicação está relacionado ao domínio técnico do tema fonte dos dados onde ambas são utilizadas, ou seja, "a especialidade na área de aplicação de dados".

Conway (2010) promove que a utilização do diagrama de Venn evidencia como a Ciência de Dados – no centro de todas as interseções – corresponde à congregação do conhecimento de matemática e estatística; habilidades de *hacker*; e sua competência significativa – entendida aqui como o contexto de aplicação. A Figura 4 expõe de maneira visual esse relacionamento.

Figura 4- Diagrama de Venn aplicado à Ciência de Dados

Fonte: adaptado de Drew Conway (2010)

Finzer (2013) detalha os conceitos que estão combinados no diagrama de maneira a construir uma sequência de cada uma das três competências no processo natural de análise por parte de um cientista de dados. Na visão do autor tem-se a seguinte ordenação:

1º) Matemática e estatística: o primeiro pensamento a ocorrer, para um cientista de dados, é o matemático - onde este é estruturado, disciplinado e quantitativo. Em seguida, o entendimento da variabilidade dos dados, bem como as ferramentas para sua análise, ocorre através da aplicação de ferramentas de estatística;

2º) Contexto de aplicação (denominado Experiência Substantiva originalmente): faz-se necessário ao cientista de dados experiência no contexto em questão, para que este seja capaz de basear a escolha de sua metodologia de abordagem, sem a qual sua abordagem pode se tornar muito mais dificultosa ou, em alguns casos, até impossível de gerar resultados plausíveis;

3º) Habilidades de computação: com a utilização de técnicas computacionais que variam entre diferentes *softwares*, somada à sua criatividade e capacidade de resolver problemas, o cientista de dados é capaz de agrupar dados de maneira que sua visualização faça sentido e seja possível se extrair informações anteriormente ocultas em meio ao seu universo.

Vale destacar, que ao se pesquisar a literatura vigente de capacitação sobre o tema, o que se encontra são preparações e técnicas que abordam justamente a primeira e terceira competências, enquanto a segunda fica a cargo da experiência adquirida através de vivência. Na visão de Grus (2016), isso se dá já que conceitos de matemática, estatística e computação são mais facilmente transmitidos de maneira estruturada – conhecimento explícito - enquanto a "competência significativa" acontece majoritariamente de maneira tácita. O que fica evidente quando o próprio autor indica em sua obra *"Data Science from Scratch"* que: "planejei escrever um livro abordando as três, mas eu rapidamente percebi que uma abordagem de "competência significativa" exigiria dezenas de milhares de páginas. Assim, decidi focar nos dois primeiros."

2.2.1 *DATA MINING*

A tradução da expressão da língua inglesa *Data Mining* para o português é mineração de dados. A priori, entender o significado das duas principais palavras do termo em separado – mineração e dados – ajuda a compreender o porquê de formarem o conceito em questão. Para o linguista Ferreira (2020), a palavra mineração remete ao efeito ou ação de explorar minério e, em conseguinte, efetuar a sua depuração (limpeza); enquanto a palavra dados é definida como informação, bases e conhecimento. A visão linguística já traz uma ideia da razão de o termo envolver a junção dessas duas palavras, porém esta definição de dados difere daquela quando avaliada sob ótica dos autores técnicos do campo de ciência de

dados. No contexto técnico, dados não podem ser entendidos como informação e conhecimento, e sim, como formadores da base para a construção de ambos. Logo, a compreensão real do conceito exige aprofundamento técnico nos dados e seus tipos.

Aggarwal (2015) explana que os dados podem se apresentar em diferentes tipos e formatos, podendo ser quantitativos, categóricos, textuais, temporais e até em orientação de grafos. Compõem duas grandes categorias: não dependentes entre si e dependentes entre si. Para os primeiros, como o próprio nome já diz, a particularidade de um dado não impacta na de um segundo (o número da casa de um indivíduo independe do nome de sua rua). Os segundos já apresentam subordinação (a existência de um grupo de pesquisa está submetida à presença de um professor orientador). Enquadrá-los na categoria correta, para o mesmo autor, é atividade fundamental para escolha correta dos algoritmos a serem utilizados, posto que a atividade de *Data Mining* está justamente baseada na busca pelo relacionamento e agrupamento de dados de modo a ser possível identificar padrões.

Atualmente existe divergência de conceituação concreta do termo *Data Mining* por parte da academia. Enquanto alguns autores como Aggarwal (2015) definem que este é o estudo de coleta, limpeza, processamento, análise e desenvolvimento de *insights* a partir de dados, ou seja, o processo completo de uma ponta à outra; outros como Bramer (2016) adotam o conceito não como o processo completo, mas como a sua parte central. A indefinição de uma caracterização única acerca do tema, entretanto, pouco impacta na sua aplicação, especialmente para o trabalho em questão, que fará uso dos dados percorrendo todas as etapas – desde a entrada até a sua saída.

Na opinião de Mendes (2018) a conquista de conhecimento demanda um processo não trivial de identificação de padrões válidos, novos e potencialmente úteis em meio aos dados. É importante salientar que ao se adotar a visão de Mendes (2018) e Bramer (2016), *Data Mining* passa a ser visto como uma das seis etapas do processo, que segundo o autor, é chamado de Descoberta de Conhecimento em Banco de Dados (DCBD). Como o trabalho em questão demanda utilização de conceitos de cada uma das etapas, estas estão descritas de maneira resumida e representadas no diagrama que se segue (Figura 5):

1) Definição do problema: remete ao diálogo entre a parte técnica e os cientistas de dados para entendimento e alinhamento do que se deseja. É importante a determinação de escopo para que a equipe tenha, mesmo que de forma rudimentar, um guia durante a análise exploratória dos dados, bem como determinação de metas a serem cumpridas;

2) Seleção de dados: a escolha de base de dados confiável e que represente a realidade do estudo em questão é fundamental no processo de legitimação de toda a análise subsequente. Não é difícil conceber que uma produção bem feita de conhecimento careça, também, de uma verificação e exploração bem feita de sua base. Grus (2016) elucida que o entendimento do contexto dos dados – que está muito além dos números apenas – impacta diretamente em interpretações estatísticas, ao exemplo de correlações enganosas sob presença de variáveis de confusão. Fato este que ocorre com relativa frequência e que só pode ser evitado através de conhecimento dos dados;
3) Eliminação de erros (limpeza dos dados): de acordo com Dalton (2019), os dados são produzidos com os mais diversos propósitos e, em geral, não foram concebidos com o intuito de serem utilizados por ciência de dados. Logo, as bases de dados, em sua maioria, apresentam problemas que precisam ser corrigidos para posterior modelagem de uma base então dita "limpa". Segundo o mesmo autor, os problemas precisam ser detectados e corrigidos antes de se dar seguimento. Na visão de Menezes (2012), a remoção de inconsistências e ruídos é o que torna a resolução possível – a título de elucidação, pode-se entender os *outliers* como exemplos de inconsistências. Em conseguinte, o último autor define como objetivo da atividade de limpeza de dados "identificar e tratar os dados que podem causar impactos negativos na mineração de dados";
4) Enriquecimento de dados: é comum o cientista de dados se deparar com bases que apresentam dados faltantes (dados *missing),* para o tal, faz-se uso de preenchimento manual (pouco utilizado no universo de *big data*) ou técnicas matemáticas, estatísticas e computacionais que se encarregam do preenchimento destes, de forma a tornar possível sua utilização. Outra prática comum é a manipulação de variáveis presentes explicitamente – de não valor aparente -, de modo a gerar novas variáveis dependentes que possuam valor futuro de análise. É interessante saber que o enriquecimento não inclui apenas adição de informação, mas também, em alguns casos, remoção. Aggarwal (2015) menciona que as bases, além de apresentarem lacunas, também podem exibir alguns dados inconsistentes que precisam ser

removidos. Vale destacar, porém, que o enriquecimento precisa ser feito de modo a garantir a manutenção de veracidade dos dados iniciais;

5) *Data Mining*: o termo mineração de dados, para Laudon (2011), remete ao processo de mergulho em um conjunto de informações que busca, através de análise, identificar as que são de interesse para produção do conhecimento desejado em questão e separá-las das que não apresentam utilidade. Atualmente este é feito de maneira computacional, devido à inviabilidade de ser feito manualmente no campo de *Big Data*. Necessário se faz, porém, informar que para alguns outros autores o termo corresponde a todas as etapas e não apenas a uma – diferente da abordagem aqui utilizada. Menezes (2012) aponta essa etapa como a essencial no processo de geração de conhecimento, já que é nela que são aplicadas as principais ferramentas inteligentes de extração e análise de dados;
6) Relatórios: de acordo com Menezes (2012), esta última etapa compreende a aplicação de técnicas de visualização de dados para exibição dos resultados obtidos através dos algoritmos e modelagens matemáticas aplicadas. Dalton (2019) salienta que a interpretação dos resultados, bem como a validação dos modelos, é parte fundamental no fechamento do processo. Os relatórios podem incluir tanto análises quantitativas quanto qualitativas, geralmente expostas através de gráficos ou simulações futuras, as chamadas análises preditivas.

Figura 5- Etapas da Descoberta de Conhecimento em Bancos de Dados

Fonte: elaborado pelo autor

Todas as seis etapas estão baseadas em aplicação de ferramentas que serão exploradas tecnicamente durante a abordagem individual de cada uma das competências descritas por Finzer (2013), nos próximos subcapítulos.

É de senso comum aos pesquisadores da área, a exemplo de Menezes (2012), que as quatro primeiras etapas da DCBD competem a uma fase chamada pré-processamento. A mesma fase merece destaque, pois segundo Mendes (2018), apesar de estar no *background* - corresponder ao suporte para a realização de *Data Mining* – esta consome cerca de 80% do tempo de trabalho do cientista de dados. Para o mesmo autor, a etapa de processamento em si – que engloba a seleção e aplicação dos modelos matemáticos de *machine learning* durante análise preditiva e utilização de ferramentas de visualização de dados, por exemplo -, acaba recebendo maior destaque por estar no final do processo, e não por toda sua demanda ao profissional. Bramer (2016) já apresenta um diferente ponto de vista, pois o autor enfatiza que apesar de os algoritmos de pré-processamento serem de grande importância, a etapa de *Data Mining* é o objeto central de todo o processo, merecendo assim, o cuidado principal.

Aggarwal (2015) elenca que existem quatro utilidades que são repetidamente encontradas dentre os vários contextos de aplicação de *Data Mining,* seja na fase de pré-processamento ou de processamento em si. Estas constituem quatro grandes blocos de fundamentação que serão estudados na subseção de *Matemática & Estatística,* porém estão exibidos como se segue: Associação de padrões*; Clustering* (agrupamento de dados); Regressão; e Classificação.

Com o intuito de facilitar o entendimento do termo mineração de dados, faz-se interessante a visualização de algumas de suas aplicações. Menezes (2012) fornece como exemplos a análise de imagens de satélite, análise de componentes orgânicos, previsão de demanda de energia, *forecasting* financeiro, sumarização de textos, análise de tendências em redes sociais e Processamento de Linguagem Natural (PLN). Segundo uma perspectiva analítica de Aggarwal (2015), exatamente pela variedade de tipos de dados, *Data Mining* é uma atividade que possui múltiplas diversidades de aplicações, o que a torna particularmente desafiadora.

Nigro (2016) se utilizou de bases referenciais estruturadas do Lattes e CNPq como fonte de dados; e os tipos de dados encontrados nessa base são majoritariamente os de orientação não-dependente: entre eles os quantitativos, categóricos e principalmente textuais. Apesar de as técnicas de mineração serem aplicáveis aos dois primeiros tipos de dados, Aggarwal (2015) elucida que sua utilização em dados de texto também ocorre, porém com

algumas modificações. Para entendimento dessas modificações e como estas se dão, faz-se necessário um campo próprio de estudo para variáveis de texto: o chamado *Text Mining*.

De acordo com Vijayarani *et al.* (2016), *Text Mining* é o processo não trivial de extração de conhecimento ou informação de dados de texto não estruturados. Em continuação, os autores indicam que a atividade compreende campos multidisciplinares como *Data Mining*, *Machine Learning*, estatística e linguística. Bramer (2016) indica que o propósito de sua aplicação é a classificação de documentos ainda não vistos, através de análise em documentos (ou objetos) pré-classificados. Entende-se que na aplicação em questão, os documentos/objetos são representados por artigos em jornais, publicações científicas completas, apenas seus *abstracts* ou até mesmo unicamente seus títulos.

Vijayarani *et al.* (2016) indicam que, assim como em *Data Mining*, o processo de *Text Mining* também necessita de uma fase de pré-processamento. Esta fase, entretanto, apresenta etapas que diferem do processo anterior justamente pela particularidade dos dados. Bramer (2016) salienta que a classificação de dados de texto, oferece uma possibilidade que era, até então, inexistente entre os outros tipos: a de múltipla classificação.

Isso quer dizer que dados de texto possuem várias categorias como Medicina, Engenharia, História, Gestão, Educação. Os autores informam a necessidade de uma etapa inicial que faça a representação de documentos de texto em dados factíveis com os processos naturais de *Data Mining*. Pode-se entender que para o processo trivial de *Data Mining* é interessante que essas categorias estejam representadas de modo binário e não categórico. A título ilustrativo, ao invés de se ter as categorias Medicina, Engenharia e História; o que se tem são respostas “sim ou não” para as perguntas “O documento é de Engenharia?”, “O documento é de Medicina?”, “O documento é de História?”.

Vijayarani *et al.* (2016) ordenam o pré-processamento de dados de texto em três etapas, enquanto Bramer (2016) - apesar de utilizar didática diferente - corrobora com o desenrolar de cada uma como se segue:

1º) *Tokenization*: esse é o primeiro passo ao se trabalhar com textos. Consiste em “quebrar” as *strings* (conteúdo textual) em pequenas estruturas ou unidades de maneira a fazerem sentido. Estas unidades são chamadas de *tokens* e podem ser entendidas como palavras, termos ou símbolos.

2º) *Stemming*: refere-se ao processo computacional de normalização (ou redução) de palavras para o seu formato primitivo ou forma base. Os algoritmos se baseiam, quase em sua totalidade, em dois métodos: o de

Porter ou Lancaster. Ambos, apesar de apresentarem comportamentos diversos, são efetivos na normalização. A título elucidativo, pode-se exemplificar que o agrupamento de palavras como “pedreiro”, “apedrejar”, “pedreira”, “pedregulho”, em uma só palavra raiz: “pedra”, é o resultado da aplicação dos algoritmos de normalização.

3º) *Stop Word Removal*: o termo inglês em questão trata da remoção das chamadas “palavras de parada”. Estas são aquelas que, apesar de serem as mais comuns em um idioma, não fornecem nenhum significado concreto ao texto e então precisam ser removidas para que seja possível extração de informação útil. Os artigos como “o”, “a”, “os”, “as”, conjunções como “e”, “então”, “para” e advérbios como “acima” ou “ao lado”, por exemplo, precisam ser removidas antes da realização de análise exploratória. Imagine que se deseje obter uma lista com as 3 primeiras palavras mais utilizadas em uma publicação científica e o resultado obtido é a seguinte lista: [“o”, “e”, “para”]. Essas três palavras não trazem informação alguma válida para desenvolvimento de conhecimento a respeito do texto. Imagine, porém, que após a remoção das *stop words*, o resultado da lista agora seja: [“dengue”, “saúde”, “prevenção”]. Já é bastante possível se ter uma ideia do assunto tratado pela publicação.

Conforme já elucidado, a utilização de *Text Mining* é bem vasta, inclusive uma de suas vertentes compreende a aplicação de seus métodos e ferramentas em documentos técnicos, o que de tão comumente praticado recebeu denominação exclusiva: *Tech Mining*. De acordo com Porter e Cunningham (2005), os algoritmos de *Tech Mining* tornam possível a extração de sentido em bases de dados textuais científicas. Segundo os mesmos autores, a exploração dessa ferramenta viabiliza novas capacidades para geração de conhecimento por parte de pesquisadores ou gestores interessados em tecnologia. Ademais, o conceito expandido também propõe a utilização das ferramentas de *text mining* em documentos de patentes, periódicos, revistas, livros e até mesmo relatórios técnicos ou dados de gestão.

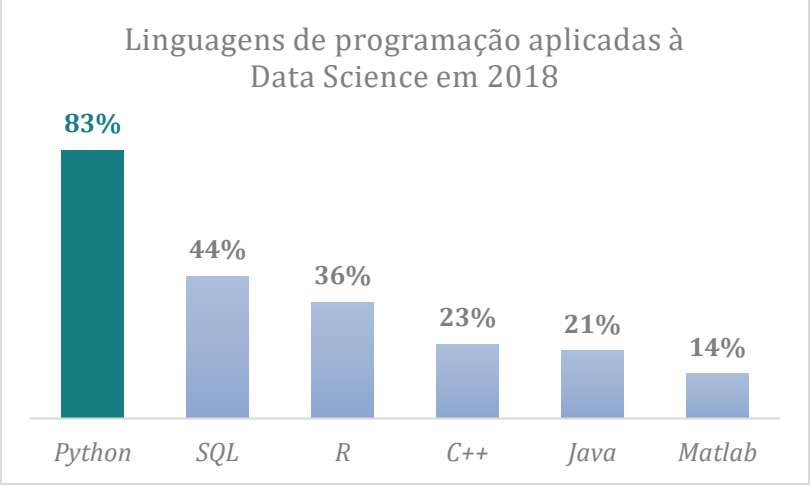
Fica evidente que todo esse trabalho de identificação de padrões em base de dados como *Data Mining* e *Text Mining*, carece de técnicas específicas até então aqui não elucidadas - já que apenas um panorama foi passado. Para Finzer (2013) compete à Ciência de Dados e suas opções em artifícios matemático-estatísticos, computacionais e particulares de cada contexto de aplicação, tornar possível a extração de conhecimento em dados de alto

volume e grande desorganização. Posto que as competências possuem ferramentas de complexa fundamentação e aplicação, o detalhamento individual a cada uma das três, de maneira a elucidá-las, ocorre nos subcapítulos conseguintes.

2.2.1.1 COMPUTAÇÃO

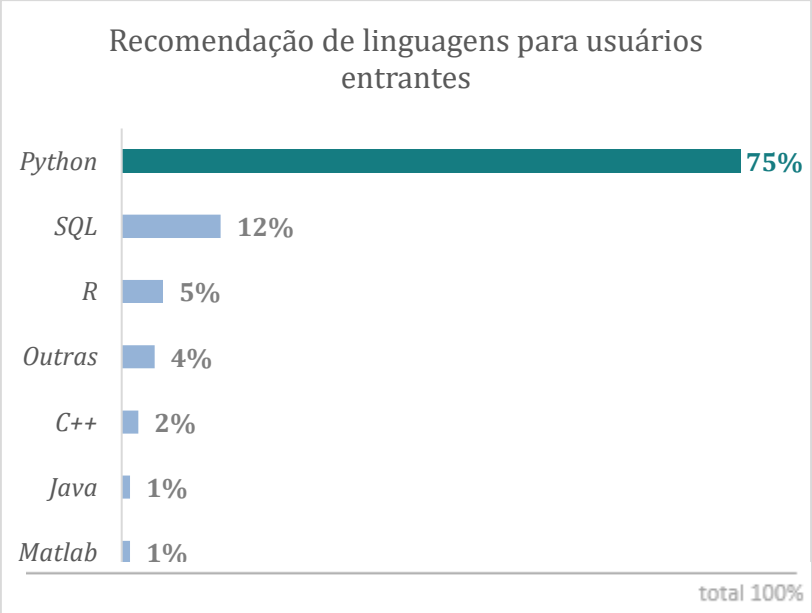
Em meio à comunidade de cientistas de dados existe um constante debate a respeito de qual a melhor linguagem de programação para se aplicar *Data Science.* Na visão de Grus (2016), Python é uma escolha evidente, apesar de um número razoável de usuários ainda considerar a linguagem de programação R e uma minoria sugerir Java ou SQL como boas opções. Em pesquisa realizada pela Kaggle (2019) – comunidade de *Data Science* do Google - foi pedido aos participantes para selecionarem todas as linguagens que utilizavam quando seu intuito era de análise de dados e destes, 83% selecionaram a linguagem Python. Em continuidade da mesma pesquisa, os usuários foram solicitados a indicar apenas uma linguagem para servir de escolha para um novo desenvolvedor entrante e Python, novamente, foi a primeira opção em maioria. Os resultados estão detalhados nas figuras 6 e 7, respectivamente.

Figura 6- Utilização de linguagens de programação em Ciência de Dados



Fonte: adaptado de Business Broadway (2019)

Figura 7- Recomendação de linguagens para ciência de dados



Fonte: adaptado de Business Broadway (2019)

As razões de Python ser a linguagem referência ao se tratar de ciência de dados são as mais diversas. Bird, Klein e Loper (2009) recomendam a utilização desta linguagem em seus trabalhos pelo fato de ser simples, mas ainda sim poderosa e com excelentes funcionalidades, especialmente quando aplicada à Processamento de Linguagem Natural. Os mesmos autores compartilham a visão de Grus (2016) quando elencam o fato de Python estar disponível para *download* em qualquer plataforma de maneira gratuita, assim como a utilização de suas bibliotecas e exibição/publicação de resultados. O último autor é ainda mais enfático ao dizer que a linguagem é de fácil codificação e é ainda mais fácil de se entender, além de apresentar grande número de bibliotecas úteis aplicáveis à *Data Science*.

Na literatura mais atual, percebe-se que Python continua sendo uma ferramenta de muito valor, dado o número de pacotes, módulos e contribuições que os *pythonistas* vêm produzindo constantemente e crescendo à uma taxa positiva (ROGEL-SALAZAR, 2020). Em resumo, não é surpresa para os autores que o número de usuários continue a crescer. Finalmente, pode-se visualizar os recursos de destaque da linguagem: gratuita - *open source;* simples de programar e de ser lida; ‘econômica” em linhas de código ao se comparar com outras linguagens; apresenta grande número de bibliotecas já disponíveis para *Data Science*; e, está em constante atualização e crescimento, já que é impulsionada por sua vasta comunidade de usuários.

Conforme evidenciado, a linguagem *Python* oferece vasta gama de pacotes e módulos (envolvendo algoritmos como os mencionados anteriormente, inclusive) já prontos e disponíveis gratuitamente. Esses pacotes são chamados de bibliotecas e podem ser entendidos como ferramentas para facilitar à execução de atividades antes complexas. Grus (2016) recomenda ao cientista de dados que, caso o interesse seja realizar algum processamento computacional, é interessante que, antes de investir o seu tempo escrevendo códigos no desenvolvimento da ferramenta que precisa, pesquise se já não existe uma biblioteca pronta capaz de realizar tal processamento; o elevado e crescente número delas indica que provavelmente ela já exista. Apenas a título elucidativo, pode-se destacar a existência de bibliotecas capazes de plotarem gráficos diversos, realizarem operações matriciais, estruturar dados para trabalhos em *dataframes*, operar os algoritmos de PLN, *machine learning*, entre outros. Cabe aqui, o destaque de algumas das bibliotecas principais relacionadas ao estudo.

a) *Pandas*: segundo os seus desenvolvedores, Mckinney e Kluyver (2020), esta biblioteca é a ferramenta adequada para trabalhos com dados armazenados em tabelas ou planilhas. Ela possuiu atributos capazes de ajudar o usuário a limpar, processar e visualizar suas bases de dados. Além de já possuir algumas ferramentas estatísticas básicas embutidas e permitir integração com outras bibliotecas que também estão aqui em destaque, como *Matplotlib* e *Numpy*.
b) *Matplotlib*: essa biblioteca permite a transformação de dados em figuras gráficas. Para Hunter *et al.* (2020), trata-se do jeito mais simples de se criar gráficos em *Python*. Além de apresentar uma vasta gama de opções de gráficos: como os de barras, linha, distribuição, entre outros; ela também permite manipulações diversas na espessura e coloração de linhas, títulos de eixos, formatos de marcadores etc. Segundo os mesmos autores, é uma opção interessante para personalização de gráficos.
c) *Numpy*: a Numpy Community (2020) define a biblioteca Numpy como um pacote para computação científica em *Python*. Essa biblioteca fornece a possibilidade de se trabalhar com vetores e matrizes multidirecionais, o que oferece rápidas operações entre elas, além de manipulações em seus formatos, ordenação, simulações aleatórias, aplicações de álgebra linear e operações estatísticas básicas. Em resumo, a comunidade *Numpy* explana

que a finalidade dessa biblioteca é facilitar e agilizar operações matemáticas avançadas (ou envolvendo números de grandes dimensões), posto que as funções já embutidas em *Python* não as fazem tão bem.

d) *NLTK*: a sigla em questão, em português, pode ser traduzida como Kit de Ferramentas para Linguagem Natural. De acordo com Bird (2009), essa biblioteca é a plataforma líder em trabalhos envolvendo PLN em *Python*. Ela fornece interfaces de fácil uso, além de oferecer algoritmos já prontos capazes de categorizar textos, analisar estruturas linguísticas e fracionar elementos gramaticais de acordo com suas funções na sentença. É considerada pela comunidade de cientista de dados como a ferramenta preferida para se trabalhar com linguagem natural, logo sua utilização é comum para a vasta maioria dos autores aqui referidos.

e) *Scikit-learn*: esta biblioteca é a mais utilizada pelos cientistas de dados que trabalham com *Python* quando seu objetivo envolve *Machine Learning*. De acordo com seus desenvolvedores, os *Scikit-learn Developers* (2017), esta ferramenta atua em situações que envolvem uma amostra de dados conhecida como entrada, para fornecer como saída previsões das características de dados desconhecidos. Os desenvolvedores afirmam que a biblioteca é capaz de lidar com problemas de aprendizagem supervisionada e não supervisionada, podendo atuar em tarefas de classificação, regressão, associação e *clustering*.

f) *Statsmodels*: seus autores, Seabold e Perktold (2010), definem a ferramenta como uma biblioteca para análises estatísticas e econometria em *Python*. Os desenvolvedores, juntamente com a comunidade de cientistas de dados, identificam o crescimento da biblioteca que se utiliza de *Python*, pelo fato de ela ser capaz de realizar funcionalidades similares à de outra linguagem de computação muito popular no meio de economistas e estatísticos: a linguagem R. Como visão geral das funcionalidades do pacote estão as regressões, modelos lineares generalizados e robustos, modelos discretos e de contraste.

Em resumo, Mendes (2018) enfatiza que compreende à ciência de dados a capacidade de se trabalhar com softwares estatísticos (como a linguagem R, por exemplo), linguagens preparadas para operar matrizes e vetores (como o Matlab) e utilização de

ferramentas tabuladas (como Excel). Entretanto, Python, por possuir todas as bibliotecas já elencadas, torna viável a utilização de apenas uma linguagem, porém com capacidade de execução combinada de todas as outras citadas.

2.2.1.2 MATEMÁTICA & ESTATÍSTICA

Da mesma forma que Aggarwal (2015) aborda duas categorias de dados – dependentes e independentes -, na visão de Bramer (2016) a nomenclatura envolve os dados rotulados e não rotulados. O que denota os primeiros é a existência de um atributo específico que os caracterize, enquanto nos últimos é impossível a definição de apenas um atributo para sua caracterização. Por conseguinte, antes de se pensar na mineração dos dados em si, é importante identificação de qual tipo se referem, pois estes são tratados de maneira radicalmente distinta em processos de *Machine Learning*.

Para os *Scikit-learn Developers* (2017), os problemas que caracterizam aprendizagem de máquina são aqueles que envolvem um conjunto de dados de entrada ou amostra e, então, procuram predizer as propriedades dos dados, inclusive daqueles desconhecidos, de saída. De maneira didática, os desenvolvedores optam por separar tais problemas em três categorias: supervisionada, não supervisionada e de reforço.

Os mesmos autores determinam que a técnica de mineração de dados que utiliza dados rotulados é definida como aprendizagem supervisionada, ou seja, é informado à máquina o que é aquela entrada de dados; enquanto a que se utiliza de dados não rotulados é chamada de aprendizagem não supervisionada, logo não ocorre essa informação. O objetivo da supervisionada é, após receber uma base com atributos bem especificados, predizer, em uma nova base, um conjunto de exemplos que ainda não foram vistos, onde estes contenham características e traços particulares já aprendidos anteriormente – o que recebe o nome de instâncias. Já o objetivo da não supervisionada é extrair o máximo de informação possível dentro da base de dados disponível (realizar uma aprendizagem) e a partir disso, replicar os padrões aprendidos em um outro conjunto de dados totalmente novo.

Vydiswaran (2017) elenca que assim como os humanos aprendem com experiências passadas, máquinas aprendem com instâncias passadas. Em resumo, na aprendizagem supervisionada é passado à máquina um padrão para o qual ela aprenda com ele e replique em novos conjuntos de dados; enquanto na não supervisionada a própria máquina identifica os padrões, para aprender com eles e os aplicar em outros grupos de dados.

Os mesmos autores citados indicam a divisão das aprendizagens em duas etapas e a explanação de cada uma, mesmo que de modo sucinto, auxilia na elucidação das diferenças entre ambas. A primeira etapa é chamada de fase de treino, enquanto a segunda é denominada de fase de inferência. As técnicas e algoritmos de treinamento e inferência variam de caso a caso e dependem, conforme já dito, dos tipos de dados – rotulados ou não rotulados. Para facilitar a diferenciação entre ambas se utiliza o Quadro 1 que se segue:

Quadro 1- Tipos de aprendizagem e suas fases

		Aprendizagem	
		Supervisionada	**Não supervisionada**
Fase	**Treino**	Dados rotulados são reunidos, indicados à máquina quais são os rótulos e seus significados e, por fim, o modelo é construído. Exemplo: é passado ao algoritmo que determinada instância é tida como positiva, enquanto outra é negativa; a partir daí o algoritmo irá aprender quais aparentam ser mais positivas do que negativas e criar um modelo para inferência futura.	Dados não-rotulados são passados à máquina, então os algoritmos iniciam um processo cíclico de aprendizagem (ou treino) onde procuram padrões e criam agrupamentos, culminando na criação de modelo para seguinte aplicação. Exemplo: são passados dados que não possuem correspondência direta com os objetivos, logo o treinamento consiste na determinação de grupos de dados similares capazes de encontrar a sua densidade de estimação.
	Inferência	Aplicação do modelo, onde o algoritmo recebe dados não-rotulados e é capaz, baseado no treino anterior, de criar os rótulos de maneira independente. Ou seja, apresenta como entrada dados não-rotulados e exibe como saída dados rotulados. Exemplo: o algoritmo recebe a instância sem classificação de positiva ou negativa, porém, após aplicação do modelo treinado, apresenta por si próprio as instâncias classificadas como negativas ou positivas.	Aplicação do modelo, onde este, seguindo a lógica que aprendeu com o treinamento da base anterior, é capaz de encontrar padrões e realizar agrupamentos em uma nova base de dados não-rotulados, gerando como saída agrupamentos e padrões não óbvios ou não evidentes aos olhos humanos. Exemplo: geração de modelos de mistura Gaussianos – um modelo probabilístico de variadas aplicações.

Fonte: elaborado pelo autor adaptado de Vydiswaran (2017)

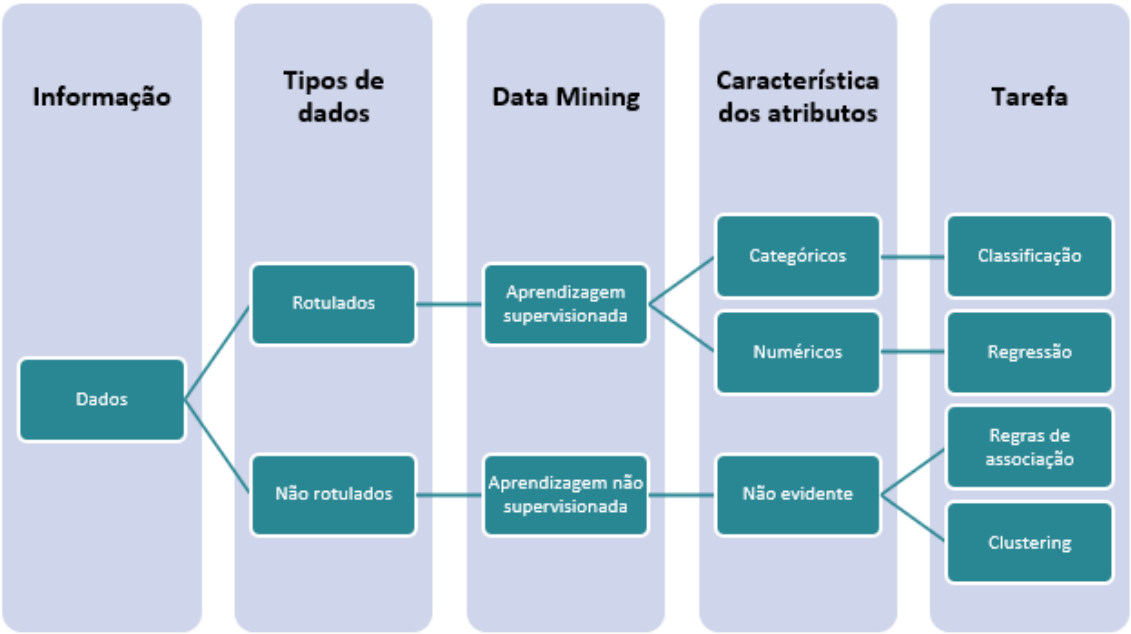
Corroborando com os autores da área, Mendes (2018) indica que se a aprendizagem supervisionada possui atributos de entrada categóricos, esta é definida como “classificação”, e caso sejam numéricos, recebe a definição de “regressão”. Um exemplo de tarefa de classificação é a identificação de imagens, por meio de algoritmos de aprendizagem de máquina, nas categorias de “cachorro” ou “gato”, onde tem-se uma classificação com rótulos

binários. Vydiswaran (2017) adiciona dizendo que se encontram também exemplos de classificações *multi-class*, ou seja, os rótulos não precisam ser apenas binários, podendo ser em números ainda maiores. Como exemplo desta classificação está a varredura de um *abstract* médico por meio de ferramentas de *Text Mining*, onde o algoritmo é capaz de identificar a que classe esse *abstract* pertence; por exemplo, ele o classifica como neurologia, entretanto, as opções iniciais de rótulos seriam pneumologia, nefrologia, dermatologia, a própria neurologia, etc.

Conforme indicado pelos autores, quando a natureza da entrada dos dados é quantitativa não faz sentido a sua saída ser em formato de categorias ou rótulos "nominais", e sim, ser uma saída numérica – a chamada regressão. Como exemplo de uma de suas tarefas, tem-se a estipulação (ou predição) do valor de um apartamento, de acordo com atributos numéricos: como o seu número de quartos, número de vagas de garagem e a sua localização – que pode ser expressa de maneira numérica através de coordenadas cartesianas ou por distâncias até determinado ponto de referência.

Bramer *et al.* (2016) fundamenta que para a aprendizagem não supervisionada, entretanto, pode-se utilizar duas diferentes abordagens: a de regras de associação ou a de *clustering*. Pode-se considerar que um algoritmo de aprendizagem de máquina que utiliza a base de vendas de uma concessionária de veículos e, automaticamente, gera categorias de compradores "esportivos" e "conservadores", é um exemplo desta abordagem. Perceba, diferentemente da tarefa de aprendizagem supervisionada definida como classificação – onde o algoritmo receberia as categorias "esportivos", "conservadores", "SUV", entre outras, em sua entrada de dados, por exemplo - o algoritmo em questão não recebe nenhuma opção prévia de categoria ou tipo de rótulo, logo, é capaz de desenvolver essa classificação através de sua própria aprendizagem. A apresentação de uma ilustração horizontal torna a visualização dos conceitos mais compreensível (Figura 8).

Figura 8- Hierarquia horizontal de *Data Mining*



Fonte: elaborado pelo autor

Em acréscimo, e a título elucidativo, a tabela a seguir indica quais são os algoritmos mais comummente utilizados em cada tipo de dados:

Quadro 2 - Aplicações mais comuns de algoritmos aos tipos de aprendizagem

Tipo de aprendizagem	Exemplo de algoritmo
Supervisionada	K-Nearest neighbors
	Naïve Bayes
	Árvore de decisão
	Regressão linear simples
	Regressão logística
	Support Vector Machine
Não supervisionada	K-Means
	Agrupamento Hierárquico *Bottom-up*

Fonte: Adaptado de Gonçalves (2012) e Grus (2016)

Grus (2016) salienta que, apesar da aplicação dos algoritmos de aprendizagem de máquina receber o destaque em si, o conhecimento de uma série de conceitos matemáticos e estatísticos que ocorrem de pano de fundo é essencial para entendimento do funcionamento dos algoritmos aplicados. Cabe, então, breve explanação sobre dois importantes modelos estatísticos, como se segue:

- Estatística descritiva: Mishra *et al*. (2019) definem este modelo de estatística como aquele capaz de descrever de maneira resumida e válida todos os tipos de informação contidas em uma base de dados utilizando apenas algumas palavras ou características estatísticas básicas como média ou desvio padrão. Admite-se que é um estudo voltado para inferências que abrangem um recorte temporal presente ou passado através de histórico de dados. Pode-se enquadrar este modelo descritivo em três grandes tipos: medidas de frequência (englobando o número absoluto de ocorrências de um evento ou seu percentual em comparação ao todo); medidas de tendência central (aquelas que fazem comparação de dois ou mais grupos com medidas estatísticas centrais "clássicas" como média, mediana ou moda; e as medidas de variação (onde é exibido o quanto os dados estão dispersos entre si, ou seja, o grau de variação dentro de uma população ou amostra, como indicado no desvio padrão, na variância e nos quartilhes, por exemplo).
- Estatística inferencial: Segundo os mesmos autores, o modelo inferencial já é aquele que desenha conclusões baseado em dados sujeitos a variações randômicas – como erros observacionais ou variação amostral. Na estatística inferencial, em sua maioria, o foco está na predição para comportamentos futuros ou na generalização sobre uma população através de estudo de uma de suas amostras menores. Moore *et al*. (2016) complementa que existem dois tipos mais comuns deste modelo: os intervalos de confiança e os testes de significância. O primeiro é utilizado quando o objetivo é estimar parâmetros de uma população, já o segundo é aplicável quando o foco está em adquirir uma evidência em favor de uma afirmação sobre os parâmetros da população.

Grus (2016) infere que o cientista de dados que apresenta conhecimento dos algoritmos destaca-se na resolução de problemas complexos e possui maior senso crítico no momento de seleção da ferramenta adequada a se utilizar. Posto que o foco computacional deste trabalho está, primordialmente, na aplicação de *Text Mining*, faz-se interessante exploração teórica de três dos seus principais algoritmos de classificação: *Naive Bayes*, Regressão Logística e *Support Vector Machine* (SVM). A priori, é importante destacar que se trata de três algoritmos de *machine learning*, logo a divisão das amostras em grupos de

treino e teste é comum a todos. Contudo, cada um apresenta particularidades no que tange à abordagem, tendo em vista que apresentam diferentes técnicas de cálculos de probabilidade, estatística, matemática e otimização. A diferenciação de cada um está nos parágrafos seguintes.

Ao se pensar em *Naive Bayes*, o primeiro ponto a se destacar, segundo Gonçalves (2002), é que este algoritmo está fundamentado na teoria *Bayesiana*, ou seja, em probabilidades incondicionais. Grus (2016) descreve que "a chave para *Naive Bayes* é fazer a suposição de que as presenças (ou ausências) de cada palavra são independentes umas das outras". Justamente por essa independência entre probabilidades ser uma hipótese extrema, a técnica recebe o nome *naive* (inocente em português).

Partindo agora para a lógica por trás do algoritmo, a autora Gonçalves (2002) descreve que *Naive Bayes* atua através de uma operação que admite que "a probabilidade de ocorrência de uma conjunção de atributos é igual ao produto das probabilidades de ocorrência de cada atributo isoladamente". De modo a elucidar o conceito de maneira prática, Vydiswaran (2017) traz o exemplo de classificação da palavra *Python*. Esta palavra pode ser enquadrada primordialmente em três classes: zoologia (por representar o nome de uma espécie de cobra - Píton), entretenimento (correspondendo a uma famosa série de comédia britânica – *Monty Python Tv Show*, ou ciência da computação (quando indica o nome da linguagem de computação interpretada – *Python*). A maneira como a máquina classifica a palavra está pautada nas probabilidades, onde tem-se, por exemplo, que a probabilidade de a classe zoologia ser atribuída à palavra *Python*, é a probabilidade de a classe ser zoologia sem nenhuma informação, multiplicado pela probabilidade de se encontrar a palavra *Python* dado que já é sabido que se trata de um documento de zoologia, dividido pela probabilidade de ocorrência da palavra *Python* sem nenhuma informação. Raciocínio análogo pode ser realizado na seleção das demais classes.

Vale destacar, que o algoritmo também investiga palavras combinadas, ou seja, Vydiswaran (2017) informa que a palavra *Python* é comumente associada à classe zoologia, porém as palavras *Python download* combinadas aumentam as chances de associação à classe ciência da computação, por exemplo.

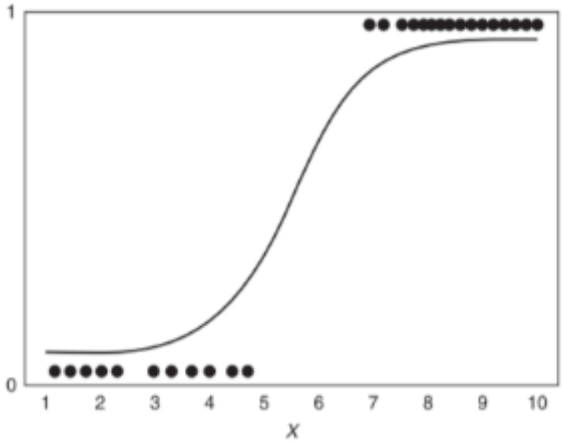
Sobre o segundo algoritmo - Regressão Logística - Hair *et al.* (2009) inferem que esse modelo de regressão, assim como a regressão linear, está pautado na relação existente entre a variável dependente e as variáveis independentes. Entretanto, na regressão logística, a variável dependente se apresenta de maneira binária, categórica ordenada ou categórica não

ordenada. A variável binária é aquela em que se detecta a presença de determinado atributo ou não (zero ou um, sim ou não), enquanto a categórica ordenada apresenta hierarquia entre as respostas (ruim, razoável, bom, ótimo) e a não ordenada corresponde a características que não possuem relações de ordem entre si (comédia, romance, terror).

Grus (2016) elucida que o propósito deste modelo é gerar saídas que façam uma previsão de pertencimento de classe, ou seja, corresponde à análise preditiva. Hair *et al.* (2009) explanam que a regressão logística se utiliza da função logística como meio de representar a relação entre a variável dependente e as independentes. Os mesmos autores, - pautando seu exemplo em uma categoria binária para facilitar o entendimento do conceito -, informam que os valores previstos obrigatoriamente permanecem entre zero e um; onde Grus (2016) descreve que "conforme sua entrada fica grande e positiva, ela se aproxima cada vez mais de um. Conforme sua entrada fica grande e negativa, se aproxima mais de zero".

O desenho dessa curva logística se utiliza de modelos como minimização de erros quadrados ou máxima verossimilhança, onde o enfoque está em maximizar a probabilidade de ocorrência de determinado evento. Grus (2016) e Hair *et al.* (2009) indicam que é necessário avaliar a "qualidade do ajuste do modelo" e, portanto, utiliza-se do chamado exame de precisão preditiva (ou matriz confusão) que resulta em um importante indicador: o R^2. Vale destacar – a título elucidativo -, que Vydiswaran (2017) admite que valores de R^2 próximos à 0.70 são indicativos de modelos de boa qualidade. A figura que se segue representa um bom ajuste probabilístico em relação às variáveis.

Figura 9- Curva logística com relação bem definida



Fonte: adaptado de Hair *et al.* (2009)

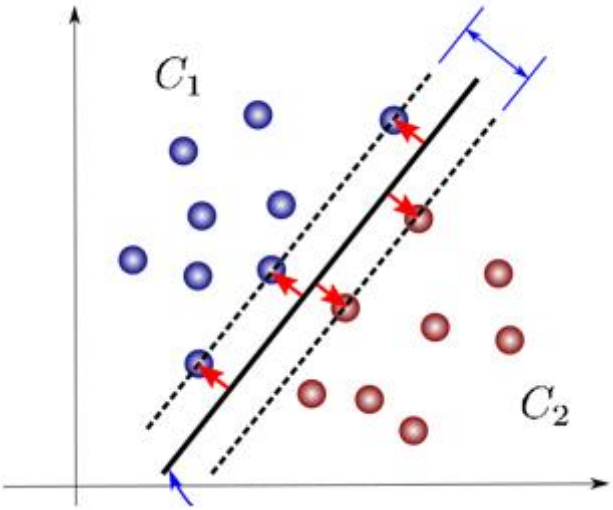
Sobre o último algoritmo - *Support Vector Machine (SVM)* - Gonçalves (2002) descreve que essa classe corresponde àqueles baseados em aprendizagem estatística resultantes de problemas de otimização quadrática. Por se tratar de controle de generalizações, utiliza-se de métodos como Princípio de Minimização do Risco Estrutural, análise discriminativa linear, *perceptron*, entre outros, e é amplamente aplicado com o intuito de identificação de padrões. A autora prossegue e descreve que o algoritmo opera de maneira a criar uma fronteira de decisão, onde essa é composta por termos de um subconjunto do grupo de dados de treinamento da aprendizagem: os vetores suporte.

Grus (2016) corrobora dizendo que essa fronteira de decisão nada mais é do que um "limite que divide o espaço de parâmetro entre duas partes de espaço correspondentes", por exemplo contendo ou não contendo determinado atributo. Vydiswaran (2017) indica que essa fronteira pode apresentar diversos formatos (linhas, quadrados, formas não definidas) e a sua determinação (que são os vetores de suporte) ocorre exatamente porque o modelo é capaz – através de técnicas de otimização - de encontrar a melhor figura geométrica que separa as classes nos dados de treinamento.

Vydiswaran (2017) adiciona, entretanto, que durante o processo de treinamento em *machine learning* erros ocorrem, então é comum encontrar eventos de um determinado parâmetro dentro da fronteira cuja finalidade seria de separá-los. Logo, os autores sugerem a aplicação de otimização para que se encontre a melhor fronteira linear – em um plano n dimensional representado por um hiperplano – já que esta é uma opção mais interessante na divisão dos dados, pois, apesar de não apresentar acurácia de 100%, ainda sim é simples de ser construída, permite fácil visualização e é o melhor que se pode conseguir. Ou seja, corrobora com o axioma de que modelos simples proporcionam boas generalizações.

Em resumo, Vydiswaran (2017) informa que os SVM são classificadores lineares capazes de encontrar essa função linear (ou hiperplano) que separa as duas classes de dados. É importante acrescentar que, apesar de os SVM trabalharem de maneira binária, eles também funcionam com classificações de mais de duas classes (*multi-class*). Para o tal, são aplicadas técnicas de comparação de uma classe com todas as demais ou de comparação de classes uma a uma. A figura 10 indica graficamente essa classificação.

Figura 10 - Representação gráfica de *Support Vector Machine*



Fonte: adaptado de Vydiswaran (2017)

2.2.1.2 CONTEXTO DE APLICAÇÃO

Grus (2016) infere que o cientista de dados carece do que chama de competência significativa. Na visão do autor a expressão compete ao domínio do contexto em que os dados estão inseridos tanto para sua extração, manipulação e aplicação. Pensando-se na identificação de um portfólio acadêmico, as informações sobre o conhecimento produzido por uma IES estão localizadas em artigos e documentos científicos. O conhecimento dessas bases de dados e dos possíveis resultados esperados após análise exploratória, compõem o seu contexto de aplicação. O acesso a eles é fornecido por diferentes entidades e seus *websites*, porém algumas merecem destaque, especialmente pela UFJF ser uma IES brasileira, como a *Web of Science* e o portal de periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

De acordo com a CAPES (2020), esta é uma fundação do ministério da educação com finalidade de promover expansão da produção científica brasileira. A fundação atua de modo a oferecer incentivos à comunidade científica, avaliações de pós-graduação, estímulos à geração e apoio à formação de professores, além de disseminação e desenvolvimento de conteúdos acadêmico-científicos, sendo os últimos agrupados em um portal na internet.

Tal biblioteca, que segundo a CAPES (2020) pode ser entendida como um acervo online, é denominada de periódicos CAPES. Esses remetem a uma biblioteca virtual que contém mais de 45 mil arquivos científicos textuais na forma de periódicos, livros e obras

em diversas bases. Ela pode ser acessada por professores, alunos e pesquisadores vinculados às suas instituições parceiras. Os periódicos também estão disponíveis de modo remoto via Comunidade Acadêmica Federada (CAFe), onde o usuário, por meio de login e senha oriundos de sua instituição, consegue acessar bases de dados mais completas e variadas – inclusive a *WoS*.

Segundo sua produtora Clarivate Analytics (2020), a *WoS* é uma ferramenta de pesquisa que permite ao usuário adquirir, analisar e disseminar informação sobre ciência oriunda de artigos e documentos científicos. Além de permitir o *download* de publicações na íntegra, o *website* também viabiliza documentos em formatos estruturados que contenham informação de diversas publicações no mesmo arquivo de modo compilado. A título de exemplificação, a ferramenta permite a saída de um documento tabulado que contenha dados de até 500 diferentes artigos, onde nesse se encontram conteúdos como os nomes completos dos autores e coautores, título da publicação, ano da publicação, nome da instituição, *abstract*, área de estudo, entre outros. Ou seja, um vasto conjunto de atributos categóricos e numéricos a ser explorado.

Como o cabeçalho das colunas é exibido de forma abreviada e em inglês, o glossário em formato de quadro a seguir auxilia na identificação de cada atributo utilizado neste estudo:

Quadro 3- Glossário das colunas presentes na base de dados da *WoS*

Ícone	Legenda
PT	Tipo de publicação
AU	Autor
AF	Nome completo do autor
TI	Título
SO	Nome do periódico
LA	Idioma
DE	Palavras-chave do autor
DI	Palavras-chave *WoS*
AB	Resumo
RP	Endereço do autor principal
C1	Endereço demais autores
TC	Citações na *WoS*
Z9	Citações totais
U1	Nº de acessos nos últimos 180 dias
U2	Nº de acessos desde 2013
PY	Ano de publicação
WC	Categoria da *WoS*
SC	Área de estudo

Fonte: elaborado pelo autor

3. DESENVOLVIMENTO

Nesta etapa do trabalho está descrito o passo a passo do que foi realizado desde o pré-processamento, passando pela análise exploratória, aplicação de ferramentas de *data mining* e *text mining* e finalização com aplicação de ferramentas de *machine learning* e exibição de *dashboard*.

A extração dos dados ocorreu na WoS e estes passaram por um filtro inicial no site onde selecionou-se a UFJF como unidade de ensino e artigos publicados do período de 2000 até 2020. Os dados foram baixados em formato txt, porém puderam ser abertos no Excel, tal qual um arquivo no formato csv.

3.1 MANIPULAÇÃO EM EXCEL

Após inicialização do software Excel, a primeira coluna foi selecionada e na aba "*Data*" clicou-se na opção "*text to columns*" e a flag "*delimited*" foi acionada. Dentro da *flag*, o delimitador escolhido foi "*semicolon*" (ou ponto e vírgula na versão em português). Com isso, o software dividiu cada coluna com seu respectivo cabeçalho e dados referentes.

O pré-processamento iniciou-se na ferramenta Excel, onde o autor identificou quais cabeçalhos seriam relevantes para análise exploratória e, deste modo, foram excluídas diversas colunas, enquanto outras foram repartidas para auxiliar na separação adequada dos dados. As colunas "Instituicao" e "Pais Autor Principal" merecem destaque, pois foram adicionadas ao arquivo original. A primeira foi obtida através da separação da coluna "Endereco Autor Principal" com a ferramenta Excel *text to columns*, onde o separador utilizado foi a vírgula e a informação utilizada estava contida na terceira coluna gerada. Para a segunda, a extração ocorreu também através da coluna "Endereco Autor Principal", entretanto, foi aplicada uma fórmula (de modo genérico *=TRIM(RIGHT(SUBSTITUTE(text," ",REPT(" ",100)),100)*) que extrai a última palavra de uma *string* e, com isso, o resultado foi o país do autor principal. Ambas colunas foram inseridas ao arquivo original após extração.

Ao cabeçalho "Categoria WC" foram adicionadas mais seis colunas (categoria 1 a 6) e ao cabeçalho "Area" foram acrescidas mais quatro (Area 1 à Area 4), de modo que houvesse uma coluna para cada categoria e área mencionada dentre os artigos. A conclusão desse processo totalizou 30 colunas, onde essas poderiam conter informações ou se apresentar vazias na ausência de dados.

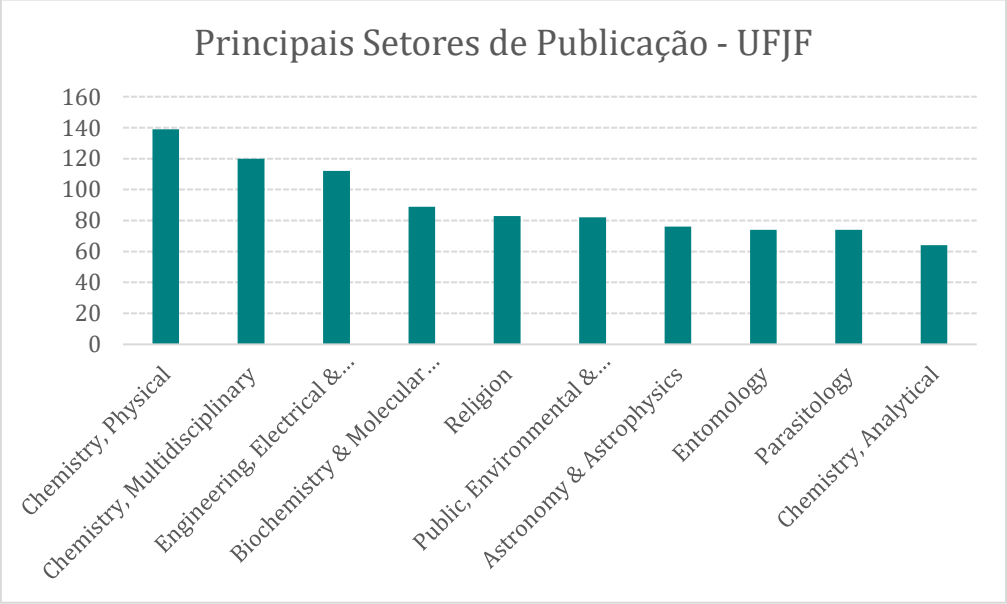
O título de cada uma delas pode ser visualizado na citação do código a seguir, onde foi possível extrair os nomes e inseri-los em uma lista no Python.

> ***In:*** *df.columns*
>
> ***Out:*** *Index(['Tipo', 'Autor', 'Autor Completo', 'Titulo', 'Periodico', 'Idioma', 'PC Autor', 'PC WC', 'Resumo', 'Endereco Autor Principal', 'Instituicao', 'Pais autor principal', 'Agencia de Fomento', 'Citacoes WC', 'Citacoes Totais', 'Acesso 180 dias', 'Acesso 2013', 'Ano', 'Categoria WC', 'Categoria1', 'Categoria2', 'Categoria3','Categoria4', 'Categoria5', 'Categoria6', 'Area', 'Area1', 'Area2', 'Area3', 'Area4'], dtype='object')*

Aproveitou-se os dados ainda na planilha de Excel para realização de análise exploratória inicial. Essa primeira análise consistiu em realizar a contagem de publicações incluídas na "Categoria 1", ou seja, a que corresponde à primeira indicada no enquadramento do artigo – a que pode ser considerada a principal. Para tal, utilizou-se de uma tabela dinâmica composta por todos os dados e que foi trabalhada por filtros, contagem de valores e linhas. Para o campo "filtros" selecionou-se todas as opções que envolviam a UFJF e suas variações como: "Juiz de Fora Univ", Univ Fed Juiz de Fora", "Fed Univ Juiz de Fora", "Fed Univ Juiz de Fora Brazil" e similares. No campo "linhas" adicionou-se "Categoria1" e no campo "Valores" selecionou-se "contagem de Categoria1".

A tabela dinâmica foi colocada em ordem decrescente de valores e exibiu 193 diferentes categorias com o número de artigos em que apareceram. As dez primeiras categorias com maior participação em número de artigos podem ser visualizadas no gráfico a seguir:

Figura 11 - Dez setores em maior número de publicações



Fonte: o autor

3.2 MANIPULAÇÃO EM PYTHON

Como este trabalho tem o enfoque em utilização de ferramentas de *Data Science, Text Mining e Data Mining* para levantamento do portfólio acadêmico da UFJF e a ferramenta escolhida para realizá-los foi *Python*, agora podem ser observadas as manipulações que se fizeram necessárias. O caminho percorrido e as etapas foram realizadas no *Jupyter Notebook* (interface *amigável* para desenvolvedores *Python*) e estarão descritas a partir daqui.

No primeiro momento, o primeiro bloco de códigos foi voltado para importação das bibliotecas que viriam a ser utilizadas e este bloco pode ser visualizado como segue:

```
import os
import subprocess
import stat
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib as mat
import matplotlib.pyplot as plt
from datetime import datetime
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
sns.set(style="white")
%matplotlib inline
import nltk
nltk.download('rslp')
nltk.download('stopwords')
nltk.download('punkt')
import re
```

Logo após, foi criado o *Dataframe* através de importação da base de dados trabalhada anteriormente em Excel e duas novas colunas foram adicionadas, onde estas estão compostas apenas por zeros ou uns. A primeira coluna tem o intuito de incluir os artigos considerados bem citados - aqueles que foram citados mais de três vezes receberam 1, o demais zero. A segunda coluna aqueles considerados bem visualizados - os periódicos que tiveram mais de dez acessos desde 2013 com o número 1 e os demais com zero. Vale destacar,

que para a construção da coluna binária utilizando filtros de três citações (ou mais) partiu-se do bom senso.

```
df = pd.read_excel('base_exploratoria_1.xlsx')
df.head()
df['bem_citados'] = np.where(df['Citacoes Totais'] > 3, 1, 0)
df['mais_vistos'] = np.where(df['Acesso 2013'] > 10, 1, 0)
```

A função *groupby* tornou possível a visualização da proporção de cada uma delas. A entrada dos códigos e suas respectivas saídas podem ser visualizadas a seguir:

```
In: df.groupby('bem_citados').size()
Out: bem_citados
     0    3261
     1    4239
     dtype: int64
In: df.groupby('mais_vistos').size()
Out: mais_vistos
     0    4596
     1    2904
     dtype: int64
```

A primeira etapa do pré-processamento para *text mining* incluiu modificar os textos inteiramente para letras minúsculas e, em seguida, remover as pontuações presentes. Tais manipulações foram aplicadas apenas às colunas cujo texto foi interessante para análise. São elas: resumo, palavras-chave definidas pelo autor e palavras-chave WC.

```
#transformação para minúsculas
df['Resumo'] = df['Resumo'].apply(lambda x: " ".join(x.lower() for x in
str(x).split()))
df['PC Autor'] = df['PC Autor'].apply(lambda x: " ".join(x.lower() for x in
str(x).split()))
df['PC WC'] = df['PC WC'].apply(lambda x: " ".join(x.lower() for x in
str(x).split()))


#remoção de pontuação
df['Resumo'] = df['Resumo'].str.replace('[^\w\s]','')
df['PC Autor'] = df['PC Autor'].str.replace('[^\w\s]','')
df['PC WC'] = df['PC WC'].str.replace('[^\w\s]','')
```

Em seguida iniciou-se o processo de limpeza de textos através da remoção das *stop words*, das palavras mais comuns, mas não muito relevantes e das que aparecem raramente e também não se apresentam relevantes de mesmo modo. No caso, optou-se pela remoção das 35 primeiras palavras da lista de comuns e não relevantes, posto que a 36ª é "*energy"* e pode ser considerada relevante.

```
# Stop Word Removal do resumo
from nltk.corpus import stopwords
stop = stopwords.words('english')
df['Resumo'] = df['Resumo'].apply(lambda x: " ".join(x for x in x.split() if
x not in stop))

#Limpeza ainda mais robusta da base identificando as 35 palavras mais
comuns e não muito relevantes
#A palavra 36 é "energy" e foi considerada relevante
mais_frequentes = pd.Series("
".join(df['Resumo']).split()).value_counts()[:35]


#Aplicando limpeza mais robusta no resumo com as palavras frequentes
mais_frequentes = list(mais_frequentes.index)
df['Resumo'] = df['Resumo'].apply(lambda x: " ".join(x for x in x.split() if
x not in mais_frequentes))

# Do mesmo modo, remover as palavras que aparecem raramente, e por
isso não têm relevância
menos_frequentes = pd.Series('
'.join(df['Resumo']).split()).value_counts()[-8850:]
# 8850 porque é a quantidade que melhor separa, nessa base, as palavras
que aparecem só uma vez, das que aparecem 2.
```

Neste momento, a limpeza textual - que pode ser considerada mais robusta - já se encerrou, com isso, deu-se início um segundo processo de limpeza, porém mais refinado. Tal processo engloba o *stemming* (remoção de sufixos e afixos) e a *tokenization* (transformação das palavras para sua forma raiz/primitiva). A título didático, o segundo processo pode ser entendido como um *stemming* com maior profundidade.

```
#Aplicando stemming: remocao de sufixos
from nltk.stem import PorterStemmer
st = PorterStemmer()
df['Resumo'].apply(lambda x: " ".join([st.stem(word) for word in x.split()]))

nltk.download('wordnet')
from nltk.tokenize import word_tokenize
from nltk.stem.wordnet import WordNetLemmatizer
lemmatizer = nltk.stem.WordNetLemmatizer()
wordnet_lemmatizer = WordNetLemmatizer()

# Lemmatizing: convercao das palavras para sua forma raiz
df['Resumo'] = df['Resumo'].apply(lambda x: "
".join([wordnet_lemmatizer.lemmatize(word) for word in
word_tokenize(x)]))
```

As manipulações ocorridas marcam o fim do pré-processamento de textos ou limpeza. As seguintes competem ao que é considerado neste trabalho, de fato, como *text*

mining. Para esta etapa estão incluídas atividades de *machine learning*, análise de acurácia dos modelos e os resultados com os termos mais empregados. Para a exibição dos resultados, a escolha foi elaborar gráficos e o desenvolvimento nuvens de palavras, onde a última foi escolhida por conta de se apresentar visualmente atrativa e de interessante criação em um script *Python*.

O trabalho atual se ateve a aplicar as ferramentas de *text mining* e "cruzar" apenas duas colunas: "Resumo" e "bem citados", entretanto, outra abordagem similar aplicada às palavras chaves também faz sentido. Ao se falar desse cruzamento em si, foram aplicadas quatro abordagens distintas de *machine learning* baseadas na revisão de literatura deste estudo. Pode-se observar quais foram e o modo em que se diferem a seguir.

1. Regressão logística com aplicação de *countvectorizer*: este método é menos rebuscado e apenas conta o número de vezes que determinada palavra aparece no documento, a partir daí, indica as palavras mais frequentes sendo incapaz de ponderar sua importância para o texto em que estava inserida.
2. Regressão logística com método TD – IDF: tal opção considera mais do que o número de ocorrências de uma palavra no texto, mas também o seu peso para o conteúdo (relevância). Isso evita exclusão de palavras raras, mas importantes.
3. *Naive Bayes*: modelo discutido anteriormente na revisão.
4. *Support Vector Machine:* opção também já discutida ao longo deste trabalho.

O intuito de aplicação de quatro opções distintas – que entregariam resultados também distintos - foi justamente identificar qual apresentaria maior acurácia e maior grau de confiabilidade. Ademais, para os testes de acurácia foram utilizados três indicadores diferentes com o propósito de incremento de confiabilidade. Os testes utilizam as próprias bases de dados que foram divididas em "treino" e "teste" durante o processo de aplicação de *machine learning* para encontrar seus valores e as especificidades de cada um podem ser visualizadas em seguida.

- Indicador AUC: identifica o nível de habilidade de um classificador binário comparando uma curva com taxa verdadeira e uma com taxa falsa;
- Indicador F1: classificador binário que avalia precisão através de comparação entre resultados positivos e falso negativos;

- Indicador *Accuracy Score*: métrica para avaliação de modelos de classificação. Testa a fração de modelos de predição que estavam corretos.

Para a aplicação de *countvectorizer* iniciou-se com a divisão dos dados em 70% de treino e 30% de teste; em seguida, converteu-se o texto em *tokens*, - momento este em que foi observado que o tamanho da amostra de palavras em português não foi suficiente para configurar a base da lista - e a regressão logística foi aplicada. Em um segundo momento, foram aplicados os testes de acurácia e exibida a nuvem de palavras. A seguir, fica evidenciado as etapas de manipulação, cálculo de acurácia e resultados através dos códigos e comentários desenvolvidos em *Python.*

```
# Divisao dos dados em treino e teste
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(df['Resumo'],
                                                    df['bem_citados'],
                                                    random_state=0)

# Aplicando CountVectorizer: converte textos em vetores com a quantidade
de tokens neles contidos (token counts)
from sklearn.feature_extraction.text import CountVectorizer
vect = CountVectorizer().fit(X_train)

#Palavras citadas mais de 100 vezes nos resumos
#Aparentemente, com a base completa, as palavras em portugues nao
tiveram volume suficiente para aparecer na lista
vect.get_feature_names()[::100]
# Conversao dos dados de treino em uma matrix.
#linhas = cada documento de texto
#colunas = tokens
#conteudo das celulas = contagem de ocorrencia de cada palavra no
documento
X_train_vectorized = vect.transform(X_train)
X_train_vectorized

# Treinamento através do modelo de regressao logistica
from sklearn.linear_model import LogisticRegression

model = LogisticRegression()
model.fit(X_train_vectorized, y_train)
```

#AUC

```
In: from sklearn.metrics import roc_auc_score
    predictions = model.predict(vect.transform(X_test))
```

```
    print('AUC: ', roc_auc_score(y_test, predictions))
Out: AUC:  0.6877301290331479


#F1
In: from sklearn.metrics import f1_score
    predictions = model.predict(vect.transform(X_test))
    print('f1: ', f1_score(y_test, predictions))
Out: f1:  0.7308595584781588

#Accuracy Score
In: from sklearn.metrics import accuracy_score
    predictions = model.predict(vect.transform(X_test))
    print('Accuracy: ', accuracy_score(y_test, predictions))
Out: Accuracy:  0.6944
```

Para o modelo em questão, o teste F1 apresentou acurácia superior à 70%, entretanto os outros dois estão abaixo (porém próximos) do mesmo percentual, o que indica que o modelo foi aceitável e pode-se prosseguir para criação da lista e nuvem de palavras.

```
In: # Conversão da matriz acima em um Array Numpy
    feature_names_1 = np.array(vect.get_feature_names())

    # Ordenação da Contagem
    sorted_coef_index = model.coef_[0].argsort()

    # 20 termos menos empregados e 20 mais empregados nos artigos mais citados

    print('Menos citados Count:\n{}\n'.format(feature_names_1[sorted_coef_index[:20]]))
    print('Mais citados Count: \n{}'.format(feature_names_1[sorted_coef_index[:-21:-1]]))
Out: Menos citados Count:
     ['2019' '139' '13' '2018' 'asymptotic' 'given' 'fa' 'artificial'
     'equilibrium' 'g1' 'larger' 'origin' 'text' 'certain' 'assumption'
     'political' 'xcc' 'simulated' 'public' 'liposome']

     Mais citados Count:
     ['ltd' 'reserved' 'dark' 'lagrangian' 'recently' 'nonminimal'
     'phenomenology' 'latter' 'analog' 'intermediate' 'oscillation' 'mic'
     'vanishing' 'discussed' '2011' 'hf' 'loop' 'phonological' 'spiritual'
     'discrete']
```

Após a criação da lista, agora o enfoque está no desenvolvimento da nuvem de palavras, bem como detalhes de sua coloração, tamanho e moldes.

```
#construindo funcao para coloracao da nuvem de palavras
def random_color_func(word=None, font_size=None, position=None, orientation=None, font_path=None, random_state=None):
   h = 30
   s = int(100.0 * 255.0 / 255.0)
   l = int(100.0 * float(random_state.randint(60, 120)) / 255.0)
```

```
    return "hsl({}, {}%, {}%)".format(h, s, l)
text1 = format(feature_names_1[sorted_coef_index[:-21:-1]])

In: # Criando e gerando a word cloud image:
wordcloud = WordCloud().generate(text1)

# Exibindo the generated image:
wordcloud = WordCloud(width = 3000, height = 2000, random_state=1,
background_color='white',          color_func=random_color_func,
collocations=False, stopwords = STOPWORDS).generate(text1)
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
Out:
```

Figura 12 - Nuvem de palavras baseada em regressão logística utilizando *countvectorizer*

Fonte: o autor

A segunda opção (método TFI – DF) segue os passos de maneira análoga à anterior, ocorrendo apenas mudança na importação e aplicação de sua biblioteca específica, ou seja, a "TF – IDF Vectorizer" como exibido na primeira linha do código em sequência.

```
from sklearn.feature_extraction.text import TfidfVectorizer

# Especificado frequência mínima de ocorrência para cada termo em cada
documento de 2.
vect2 = TfidfVectorizer(min_df=2).fit(X_train)
len(vect2.get_feature_names())

In: #AUC
    X_train_vectorized = vect2.transform(X_train)
    model = LogisticRegression()
    model.fit(X_train_vectorized, y_train)
    predictions = model.predict(vect2.transform(X_test))
   print('AUC: ', roc_auc_score(y_test, predictions))
```

Out: *AUC: 0.6747693538229576*

In: *#F1*
from sklearn.metrics import f1_score
predictions = model.predict(vect2.transform(X_test))
print('f1: ', f1_score(y_test, predictions, average='binary'))
Out: *f1: 0.744736842105263*

In: *#Accuracy*
from sklearn.metrics import accuracy_score
predictions = model.predict(vect2.transform(X_test))
print('Accuracy: ', accuracy_score(y_test, predictions))
Out:
Accuracy: 0.6896

Conforme ocorrido no modelo anterior, o teste F1 apresentou acurácia superior aos demais e estes se aproximaram dos 70%, o que valida a criação da lista de palavras.

In: *feature_names_2 = np.array(vect2.get_feature_names())*
sorted_tfidf_index = X_train_vectorized.max(0).toarray()[0].argsort()
print('Menos Citados TFIDF:\n{}\n'.format(feature_names_2[sorted_tfidf_index[:20]]))
print('Mais citados TFIDF: \n{}'.format(feature_names_2[sorted_tfidf_index[:-21:-1]]))
Out: *Menos Citados TFIDF:*
['pour' 'donnees' 'desarrollo' 'parte' 'datos' 'fev1fvc' '997' 'subnational' '45809536407' '61505458122' '45740' '12443' 'straw' 'mimicked' 'cyanidin' 'qigong' 'harness' 'mcfarland' 'p0008' 'johrei']

Mais citados TFIDF:
['nan' 'captopril' 'tinnitus' 'tau' 'pinus' 'dapsone' 'jpsi' 'emphysema' 'mnk' 'juice' 'alternans' 'baptist' 'logistics' 'nc' 'clavanin' 'xylitol' 'epiphyte' 'curcumin' 'lb' 'bone']

Neste momento ocorre a criação da nuvem novamente.

text2 = format(feature_names_2[sorted_tfidf_index[:-21:-1]])
In: *# Create and generate a word cloud image:*
wordcloud = WordCloud().generate(text2)

 # Display the generated image:
wordcloud = WordCloud(width = 3000, height = 2000, random_state=1, background_color='white', color_func=random_color_func, collocations=False, stopwords = STOPWORDS).generate(text2)
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
Out:

Figura 13 - Nuvem de palavras baseada em regressão logística utilizando TF - IDF

Fonte: o autor

O terceiro modelo utilizado foi *Naive Bayes* e, ao se pensar apenas nos códigos utilizados, as diferenças são mínimas comparado aos modelos anteriores e podem ser visualizadas em cada linha de código exibida a seguir.

```
from sklearn.naive_bayes import MultinomialNB
#Gerando o modelo (Naive Bayes)
vect = CountVectorizer(ngram_range=(1,2))
freq_resumo = vect.fit_transform(X_train)
modelo = MultinomialNB()
modelo.fit(freq_resumo,y_train)

In: #AUC
    from sklearn.metrics import roc_auc_score
    predictions = modelo.predict(vect.transform(X_test))
    print('AUC: ', roc_auc_score(y_test, predictions))
Out: AUC:  0.6910869675141698

In: #F1
    from sklearn.metrics import f1_score
    predictions = modelo.predict(vect.transform(X_test))
     print('f1: ', f1_score(y_test, predictions))
Out: f1:  0.7530040053404539

In: #Accuracy
   from sklearn.metrics import accuracy_score
   predictions = modelo.predict(vect.transform(X_test))
   print('Accuracy: ', accuracy_score(y_test, predictions))
Out: Accuracy:  0.704
```

Novamente o mesmo padrão de valores de acurácia se repete e o modelo pode ser validado e partir para a elaboração da lista.

```
vect.get_feature_names()[::100]
# Conversão da matriz acima em um Array Numpy
feature_names_3 = np.array(vect.get_feature_names())

# Ordenação da Contagem
sorted_coef_index = modelo.coef_[0].argsort()

# 20 termos menos empregados e 20 mais empregados nos artigos mais citados

print('Menos citados Count:\n{}\n'.format(feature_names_3[sorted_coef_index[:20]]))
print('Mais citados Count: \n{}'.format(feature_names_3[sorted_coef_index[:-21:-1]]))
text3 = format(feature_names_3[sorted_coef_index[:-21:-1]])
```

De modo análogo é criada a nuvem de palavras.

```
In: # Create and generate a word cloud image:
    wordcloud = WordCloud().generate(text3)

    # Display the generated image:
    wordcloud = WordCloud(width = 3000, height = 2000, random_state=1, background_color='white', color_func=random_color_func, collocations=False, stopwords = STOPWORDS).generate(text3)
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis("off")
    plt.show()
Out:
```

Figura 14 – Nuvem de palavras baseada em Naive Bayes

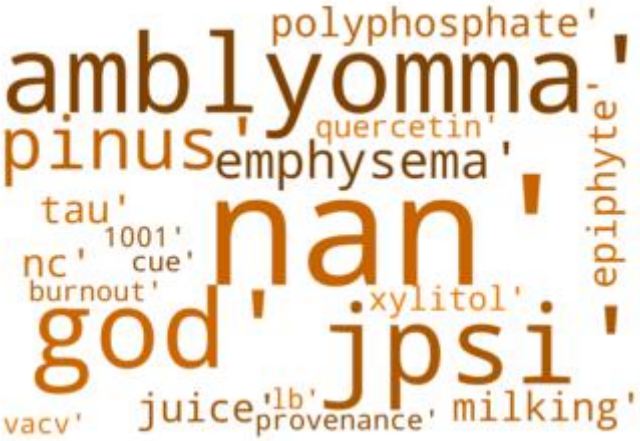
Fonte: o autor

Finalizando, a última opção foi aplicar SVM, avaliar a acurácia para validar o modelo e criar a nuvem de palavras. O código com a maneira de fazê-los está descrito em seguida. Vale destacar, que ao se pensar na elaboração da nuvem de palavras, existem múltiplas variações de paletas de cores e posicionamentos a serem aplicados, porem optou-se pela utilização de tons amarelados e marrons no estudo em questão.

```
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import model_selection, naive_bayes, svm
from sklearn.metrics import accuracy_score

#Para utilizar SVM precisamos converter textos em numeros
Encoder = LabelEncoder()
Train_Y = Encoder.fit_transform(y_train)
Test_Y = Encoder.fit_transform(y_test)

#Aplicando o metodo TD - IDF
Tfidf_vect = TfidfVectorizer(max_features=5000)
Tfidf_vect.fit(df['Resumo'])
Train_X_Tfidf = Tfidf_vect.transform(X_train)
Test_X_Tfidf = Tfidf_vect.transform(X_test)

In: # Testando SVM
    # fit the training dataset on the classifier
    SVM = svm.SVC(C=1.0, kernel='linear', degree=3, gamma='auto')
    SVM.fit(Train_X_Tfidf,Train_Y)
    # predict the labels on validation dataset
    predictions_SVM = SVM.predict(Test_X_Tfidf)
    # Use accuracy_score function to get the accuracy
    print("SVM Accuracy Score -> ",accuracy_score(predictions_SVM,
Test_Y)*100)
Out: SVM Accuracy Score -> 69.12

feature_names = np.array(Tfidf_vect.get_feature_names())
sorted_tfidf_index = Train_X_Tfidf.max(0).toarray()[0].argsort()
print('Menos Citados
TFIDF:\n{}\n'.format(feature_names[sorted_tfidf_index[:20]]))
print('Mais citados TFIDF:
\n{}'.format(feature_names[sorted_tfidf_index[:-21:-1]]))
text4 = format(feature_names[sorted_tfidf_index[:-21:-1]])

In: # Create and generate a word cloud image:
    wordcloud = WordCloud().generate(text4)

    # Display the generated image:
    wordcloud = WordCloud(width = 3000, height = 2000, random_state=1,
background_color='white', color_func=random_color_func,
collocations=False, stopwords = STOPWORDS).generate(text4)
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis("off")
```

plt.show()
Out:

Figura 15 – Nuvem de palavras baseada SVM



Fonte: o autor

4. ANÁLISE DOS RESULTADOS

Cabe destacar dois vieses de análise: um direcionado para o rendimento dos algoritmos aplicados e outro voltado para a entrega do estudo proposto como um todo. O primeiro compreende a acurácia dos modelos e *insights* das nuvens de palavras, enquanto o segundo aborda as contribuições das áreas de pesquisa e seus rumos.

Através dos valores obtidos pelos testes de acurácia (AUC, F1, e *Accuracy Score)*, foi possível perceber um padrão. Todos os quatro modelos se apresentaram próximos à 70% e apontaram o teste F1 com percentual mais elevado. O que indica que um classificador binário talvez tenha melhor *fit* para modelos que contenham parte de seus dados também binários (como no caso da coluna “bem citados”). Outro ponto de destaque está na comparação das nuvens de palavras, posto que as desenvolvidas a partir de TD – IDF e SVM apresentaram palavras com significados mais específicos, enquanto as formuladas por *Naive Bayes* e Regressão Logística apresentaram palavras com sentido mais genérico.

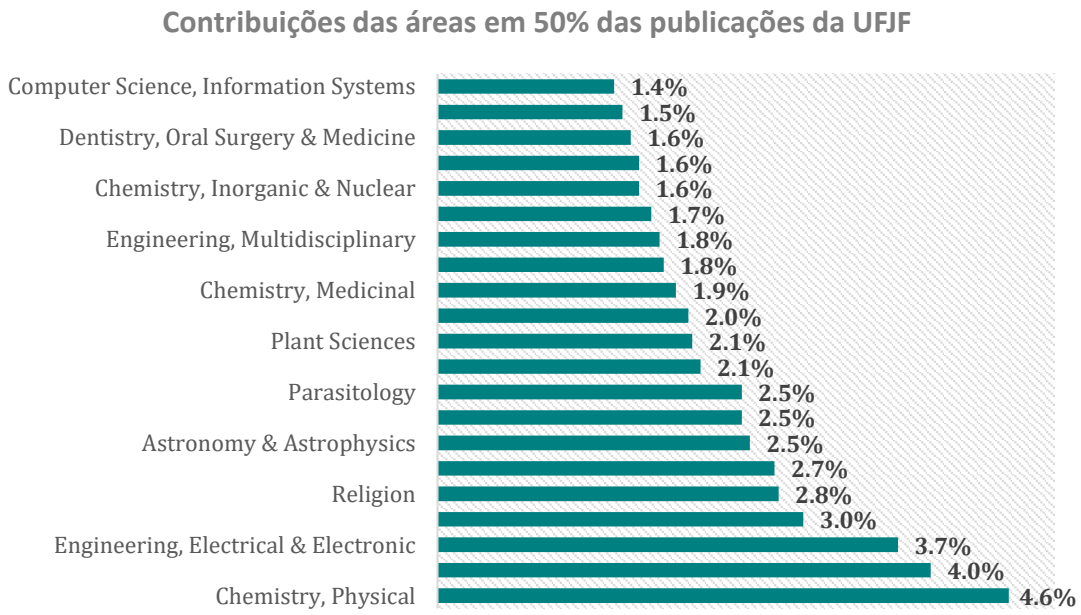
A aplicação de *text mining* e *machine learning* permitiu os algoritmos de classificação apontarem listas de palavras relevantes com acurácia média superior a 70% - o que pode ser considerado como satisfatório. Tais listas formaram as nuvens de palavras exibidas anteriormente neste trabalho e apresentam informações que podem ser interessantes para a instituição, já que permitem indicar, de modo pragmático, tópicos que têm sido mais

pesquisados. Apenas a título ilustrativo, o destaque de palavras como “boson” e “energy” representa a participação do departamento de Física como principal contribuinte para a instituição, enquanto o aparecimento da palavra “god” corrobora com o fato de a área de estudos da religião estar entre as principais na UFJF também, por exemplo.

Partindo para uma abordagem holística do estudo, a análise exploratória dos artigos da WoS extraídos ao longo dos últimos vinte anos, e que competem à UFJF, indica que os departamentos que mais contribuíram com publicações foram os de Física, Química e Engenharia Elétrica. Além disso, foi possível perceber que existe uma distribuição de contribuição razoavelmente proporcional dos demais departamentos para a instituição como um todo. Vinte e duas áreas (ou suas subdivisões) foram responsáveis por aproximadamente 50% das publicações, estando a outra metade distribuída entre 172 diferentes subdivisões.

A contribuição das áreas de Física, Química e relacionadas corresponde a menos de 10% do montante total, o que indica uma inclinação da instituição para esses vieses, mas permite inferir que a UFJF também tem mais de 90% de seus estudos pulverizados em outras. O gráfico que se segue exibe esses dados:

Figura 16- Principais áreas e seu percentual de participação



Fonte: o autor

O comparativo do gráfico percentual acima e do anterior envolvendo os dados absolutos em número de publicações possibilita entendimento de que os três principais

setores de publicações (em volume), apesar de terem participação considerável na IES, ainda somam menos de 12% da totalidade dos artigos publicados, o que indica balanceamento das contribuições das múltiplas áreas da UFJF.

5. CONCLUSÕES

Entende-se que, de maneira inicial, o trabalho foi capaz de trazer luz a algumas de suas questões propostas ao se avaliar pela ótica de pesquisa e volume de publicações. Quando se pensa em áreas em que a UFJF seja referência, fica evidente a importância do departamento de Física, Química e Engenharia Elétrica. Tal fato já era sabido de modo intrínseco e ficou evidenciado pelo número de publicações.

Ficou evidente, também, que o pré-processamento e limpeza de dados, apesar de ser eficiente, ainda exibiu a necessidade de refino posterior, posto que as nuvens de palavras indicaram palavras que foram "cortadas" ou que trouxeram erros, ao exemplo da aparição da expressão "nan". A oportunidade de tal correção evidencia o que foi discutido ao longo da revisão literária, onde foi indicado que a fase de pré-processamento compõe, em média, 80% do trabalho com dados.

Pela visualização das principais palavras classificadas através de *machine learning*, conseguiu-se inferir que a instituição vem publicando temas voltados para Física Quântica, Nutrição, Religiosidade, Medicina e Engenharia como um todo. Uma análise minuciosa das listas, permite identificar com profundidade os subcampos e entender com ainda mais precisão a direção do caminho trilhado. Palavras como "xylitol", "mass" e "logistics" já são capazes de dar um norte para identificação do rumo das pesquisas, entretanto ainda sim são consideravelmente abrangentes.

O presente trabalho deixa múltiplas portas abertas, posto que toda aplicação de *machine learning* se deu apenas cruzando a coluna "Resumo" e "Bem_citados", porém existe um campo vasto para se explorar outras colunas como "Titulo", "Periodico", "Categoria" e outras, o que permite criação de novas nuvens de palavras e indicativos do cenário de pesquisa da UFJF. Ademais, a limpeza e aplicação de algoritmos de classificação em colunas como "Endereco Autor Principal" oferece a possiblidade de exibição dos departamentos contribuintes com as pesquisas e seus relacionamentos com os demais, o que oferece a possiblidade de criação da configuração de redes de colaboração e autorias no âmbito interno e externo, resgatando informações valiosas para gestão e governança de conhecimento.

REFERÊNCIAS

ACKOFF, R. L. From data to wisdom. **Journal Of Applied Systems Analyses**, 16, p. 19-30, 1989.

AGGARWAL, C. **Data Mining**: the textbook. Suiça: Springer, 2015.

ALVES, A; BARBOSA, R. Influences and barriers to information sharing: A theoretical perspective. **Ciencia da Informação**, 2010.

BIRD, S; KLEIN, E; LOPER, E. **Natural Language Processing with Python**. Sebastopol: O' Reilly, 2009.

BRAMER, M. **Principles of Data Mining**. 3. ed. Portsmouth: Springer, 2016.

BURLAMAQUI, L. Knowledge Governance: An Analytical Approach and its Policy Implications. In: Leonardo Burlamaqui, Ana Celia Castro, Rainer Kattel. (Org.). **Knowledge Governance - Reasserting the Public Interest**. 1ed.Londres: Anthem Press, v. 1, p. 3-27, 2011.

CAPES. Periodicos CAPES. *In*: **Periodicos CAPES**: Intitu. Brasil, 2020. Disponível em: https://www.periodicos.capes.gov.br/index.php?option=com_pinstitucional&Itemid=104. Acesso em: 3 jul. 2020.

CCMD - Centro canadense para o desenvolvimento da gestão. Canadá: organizações de aprendizagem no setor público. **Revista do Serviço Público**, v. 52, n. 3, p. 100-121, 2001. Disponível em: < https://revista.enap.gov.br/index.php/RSP/article/view/311/317>. Acesso em nov. 2018.

CONAWAY, D. The data science venn diagram. 2010. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram> . Acesso em: 19 mai. 2020.

CURTY, R. G.; SERAFIM, J. S. A formação em ciência de dados: uma análise preliminar do panorama estadunidense. **Informação & Informação**, [s.l.], v. 21, n. 2, p. 307, 20 dez. 2016.

EISENHARDT, K; MARTIN, J. Dynamic Capabilities: What are they? **Strategic Management Journal.** California, p. 1105-1121, 2000.

FERRAZ, R; QUONIAM, L. Extração e Disponibilização *online* de Indicadores de Desempenho e Prospecção dos Resultados das Pesquisas em Dengue realizadas pela Comunidade Científica Brasileira por meio da Utilização da Ferramenta Computacional Scriptlattes. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação.** 2015.

FERREIRA, A. **Dicio**: dicionário online de português. Dicionário online de português. 2020. Disponível em: https://www.dicio.com.br/mineracao/. Acesso em: 08 jun. 2020.

FINZER, W. The data science education dilemma. **Technology Innovations in Statistics Education**, Califórnia, v. 7, n. 2, 2013.

FREIRE, P. *et al*. Governança do Conhecimento (GovC): o estado da arte sobre o termo: O estado da arte sobre o termo. **Biblios**, Santa Catarina, p. 21-40, 2017.

FRESNEDA, P. S. V.; GOULART, S. M. G. **A experiência brasileira na formulação de uma proposta de política de gestão do conhecimento para a administração pública federal**. Câmara dos Deputados, Brasília, 83 p., 2007.

FONSECA, M.P. **A gestão e a governança do conhecimento em Sistemas Regionais de Inovação: o caso da Zona da Mata mineira sob a ótica das instituições.** 2019. 318 f. Tese (Doutorado em Propriedade Intelectual e Inovação) – Instituto Nacional da Propriedade Industrial, Rio de Janeiro, 2019.

FOSS, N. J. **'The emerging knowledge governance approach: Challenges and characteristics**,' Organization 14: 29–52, 2007.

FOSS, N. J. Knowledge and organization in the theory of the multinational corporation: some foundational issues. **Journal of Management & Governance**, v. 10, n. 1, p. 3-20, 2006.

GRANDORI, A. FURLOTTI, M. Facio ut facias: Associational contracts for innovation. **International Studies of Management and Organization**, 2009.

GRANDORI, A. Poliarchic Governance and the Growth of Knowledge. **Knowledge Governance: Processes and Perspectives**, 2009

GRUS, J. **Data Science do Zero**: primeiras regras com python. Rio de Janeiro: Alta Books, 2016.

GONÇALVES, L.S.M. **Categorização em Text Mining**. 2002. Dissertação (Mestrado em Ciência da Computação) - USP - São Carlos, [*S. l.*], 2002.

HABERKORN, E. **Um bate-papo sobre TI: tudo que você gostaria de saber sobre o ERP e tecnologia da informação, mas ficaria encabulado de perguntar.** Itaquera: Saraiva, 2015.

HAIR, J. F. *et al*. **Análise Multivariada de Dados**. 6a ed. São Paulo: Bookman, 2009.

HAYES, B. E. **Business Broadway**. Disponível em: https://businessoverbroadway.com/. Acesso em: 19 maio 2020.

HUNTER, J. *et al*. **Matplotlib**. 2020. Disponível em: https://matplotlib.org/3.2.2/Matplotlib.pdf. Acesso em: 30 jun. 2020.

IBPAD - Instituto Brasileiro de Pesquisa e Análise de Dados. Disponível em: https://www.ibpad.com.br/blog/politica/seminario-ciencia-de-dados-e-relacoes-institucionais-demonstra-evolucao-da-forma-de-utilizar-os-dados-na-area/. Acesso em: 18 maio 2020.

KAGGLE. **Kaggle Survey 2019**. 2019. Disponível em: https://www.kaggle.com/c/kaggle-survey-2019. Acesso em: 19 maio 2020.

KLEIS, L. *et al.* Information Technology and Intangible Output: The Impact of IT Investment on Innovation Productivity**. Information Systems Research**, 23(1), 42–59, 2012.

LAUDON, K. **Fundamentos da inteligência de negócios: gerenciamento da informação e bancos de dados.** (Vol. 1 – 9º edição), 2011.

LEITE, F. C. L.; COSTA, S. M. D. Gestão do conhecimento científico: proposta de um modelo conceitual com base em processos de comunicação científica. **Ciência da Informação**, Brasília, v. 36, n. 1, p. 92–107, 2007.

LOUKIDES, M. **What is data science?** Sebastopol, CA: O'Reilly Media, 2011.

LOZADA, C.; ARAUJO, M.; GUZZO, M. Educar pela pesquisa e os museus de ciência: um estudo de caso na nanoaventura. **Revista UFG**. 2006.

MARTINS, D. L. **Introdução à Ciência de Dados para Biblioteconomia**: introdução à ciência de dados - aula 2. Brasília: Faculdade da Ciência da Informação - Unb, 2019.

MCKINNEY, W.; KLUYVER, T. **Pandas**: powerful python data analysis toolkit. Powerful Python Data Analysis Toolkit. 2020. Disponível em: https://pandas.pydata.org/docs/pandas.pdf. Acesso em: 30 jun. 2020.

MCLEVEY, J.; MCLLROY-YOUNG, R. Introducing metaknowledge: software for computational research in information science, network analysis, and science of science. **Journal Of Informetrics.** [s.l], p. 176-197, 2016.

MENDES, D. **Python Fundamentos para Análise de Dados**. Roteiro: Daniel Mendes. Brasília: Data Science Academy, 2018.

MENEZES, V.S.A. **ANÁLISE DE REDES SOCIAIS CIENTÍFICAS**. 2012. 222 f. Tese (Doutorado) - Curso de Engenharia de Sistemas e Computação, Coppe, Rio de Janeiro, 2012.

MISHRA, P. *et al*. Descriptive Statistics and Normality Tests for Statistical Data. **Annals of Cardiac Anaesthesia**, [*S. l.*], p. 67-72, 2019.

MOORE, D. *et al*. **Practice of Statistics for Business and Economics**. 4. ed. [*S. l.*: *s. n.*], 2016.

MURRAY, J. E.; SUMMER, B. E. A Revised Knowledge Pyramid. **International Journal Of Knowledge Management.** p. 19-30, 2018.

NIGRO, C. A. **Uso das ferramentas computacionais Scriptlattes, Scripgp, Patent2net para análise da produção bibliográfica e tecnológica sobre a dengue**. 2016. 152 f. Dissertação (Mestrado) - Curso de Administração, Sistemas de Saúde, Uninove, São Paulo, 2016.

NUMPY COMMUNITY (org.). **Numpy User Guide**. 2020. Disponível em: https://numpy.org/doc/stable/numpy-user.pdf. Acesso em: 30 jun. 2020.

PORTER, L.; CUNNINGHAM, W. **Tech Mining: Exploiting New Technologies for Competitive Advantage**. 2005.

ROGEL-SALAZAR, J. **Advanced Data Science and Analytics with Python**. Oxford: Crc Press, 2020.

SCIKIT-LEARN DEVELOPERS (org.). **Scikit-learn**: user guide. User guide. 2017. Disponível em: https://scikit-learn.org/0.18/_downloads/scikit-learn-docs.pdf. Acesso em: 01 jul. 2020.

SEABOLD, S; PERKTOLD, J. **Statsmodels: Econometric and Statistical Modeling with Python.** 9th Python in Science Conference, 2010.

SHARDA, R.; DELEN, D.; TURBAN, E. **Business Intelligence e Análise de Dados para Gestão de Negócios**. 4. ed. Porto Alegre: Bookman, 2019.

VIJAYARANI, S.; JANANI, R. Text mining: open source tokenization tools: an analysis. **Advanced Computational Intelligence: An International Journal.** Coimbatore, p. 37-47. 2016.

STANTON, J. *et al*. Interdisciplinary data science education. **Special Issues in Data Management**. Washington, DC: American Chemical Society, p. 97-113, 2012.

SOUZA, M.; MORAES, A.S. **Origem e evolução do conhecimento – OEC** (Vol. 1), 2012.

STOVER, W. J. **Information Technology in the Third World: can I.T. lead to humane national development?.** New York: Routledge, 2018.

TAKEUCHI, H.; NONAKA, I. **Gestão do conhecimento.** Porto Alegre: Bookman, 2008.

PELTOKORPI, V., TSUYUKI, E. **Organizational governance in internal hybrids: A case study of Maekawa Manufacturing Ltd. Corporate Governance**, 2007.

XIA, F. *et al*. Internet of things. **International Journal Of Communication Systems.** Beijing, p. 1101-1012, 2012.

WARD, S.; BARKER, A. **Undefined by data**: a survey of big data definitions. 2013.

ANEXO E – TERMO DE AUTENTICIDADE

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA

Termo de Declaração de Autenticidade de Autoria

Declaro, sob as penas da lei e para os devidos fins, junto à Universidade Federal de Juiz de Fora, que meu Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Produção é original, de minha única e exclusiva autoria. E não se trata de cópia integral ou parcial de textos e trabalhos de autoria de outrem, seja em formato de papel, eletrônico, digital, áudio-visual ou qualquer outro meio.
Declaro ainda ter total conhecimento e compreensão do que é considerado plágio, não apenas a cópia integral do trabalho, mas também de parte dele, inclusive de artigos e/ou parágrafos, sem citação do autor ou de sua fonte.
Declaro, por fim, ter total conhecimento e compreensão das punições decorrentes da prática de plágio, através das sanções civis previstas na lei do direito autoral[1] e criminais previstas no Código Penal[2], além das cominações administrativas e acadêmicas que poderão resultar em reprovação no Trabalho de Conclusão de Curso.

Juiz de Fora, 15 de março de 2021.

Miguel Nascimento Almeida
NOME LEGÍVEL DO ALUNO (A)

201749045
Matrícula

ASSINATURA

100378136-50
CPF

[1] LEI N° 9.610, DE 19 DE FEVEREIRO DE 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências.
[2] Art. 184. Violar direitos de autor e os que lhe são conexos: Pena - detenção, de 3 (três) meses a 1 (um) ano, ou multa.